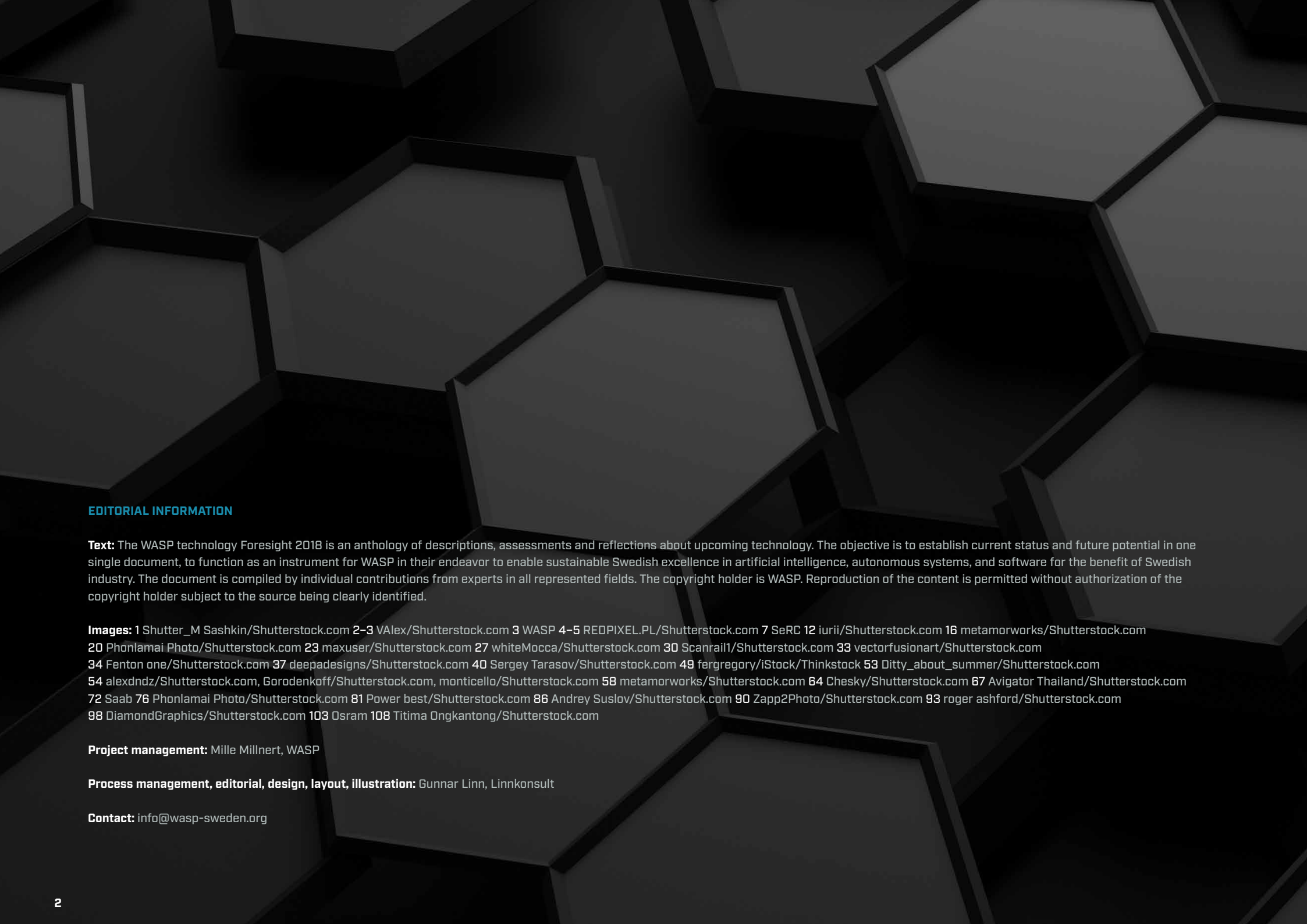
TECHNOLOGY FORESIGHT 2018

WASP's anthological assessment of upcoming technologies and their relevance for Swedish research and development

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM



EDITORIAL INFORMATION

Text: The WASP technology Foresight 2018 is an anthology of descriptions, assessments and reflections about upcoming technology. The objective is to establish current status and future potential in one single document, to function as an instrument for WASP in their endeavor to enable sustainable Swedish excellence in artificial intelligence, autonomous systems, and software for the benefit of Swedish industry. The document is compiled by individual contributions from experts in all represented fields. The copyright holder is WASP. Reproduction of the content is permitted without authorization of the copyright holder subject to the source being clearly identified.

Images: 1 Shutter_M Sashkin/Shutterstock.com 2–3 VAlex/Shutterstock.com 3 WASP 4–5 REDPIXEL.PL/Shutterstock.com 7 SeRC 12 iurii/Shutterstock.com 16 metamorworks/Shutterstock.com 20 Phonlamai Photo/Shutterstock.com 23 maxuser/Shutterstock.com 27 whiteMocca/Shutterstock.com 30 Scanrail1/Shutterstock.com 33 vectorfusionart/Shutterstock.com 34 Fenton one/Shutterstock.com 37 deepadesigns/Shutterstock.com 40 Sergey Tarasov/Shutterstock.com 49 fergregory/iStock/Thinkstock 53 Ditty_about_summer/Shutterstock.com 54 alexdndz/Shutterstock.com, Gorodenkoff/Shutterstock.com, monticello/Shutterstock.com 58 metamorworks/Shutterstock.com 64 Chesky/Shutterstock.com 67 Avigator Thailand/Shutterstock.com 72 Saab 76 Phonlamai Photo/Shutterstock.com 81 Power best/Shutterstock.com 86 Andrey Suslov/Shutterstock.com 90 Zapp2Photo/Shutterstock.com 93 roger ashford/Shutterstock.com 98 DiamondGraphics/Shutterstock.com 103 Osram 108 Titima Ongkantong/Shutterstock.com

Project management: Mille Millnert, WASP

Process management, editorial, design, layout, illustration: Gunnar Linn, Linnkonsult

Contact: info@wasp-sweden.org

WASP looks into the chrystal ball

The Wallenberg AI, Autonomous Systems and Software Program (WASP) is Sweden's largest ever individual research program, a major national initiative for strategically motivated basic research, education and faculty recruitment.

The program is initiated and generously funded by the Knut and Alice Wallenberg Foundation (KAW) with 2.6 billion SEK. In addition to this, the program receives support from collaborating industry and from participating universities to form a total budget of 3.5 billion SEK.

The vision of WASP is excellent research and competence in artificial intelligence, autonomous systems, and software for the benefit of Swedish industry. Further information about the program can be found at www.wasp-sweden.org.

As an instrument to reach the program's strategic vision, the program board has initiated the production of the forelying document, which constitutes a map of future challenges and possibilities within the scope of WASP.

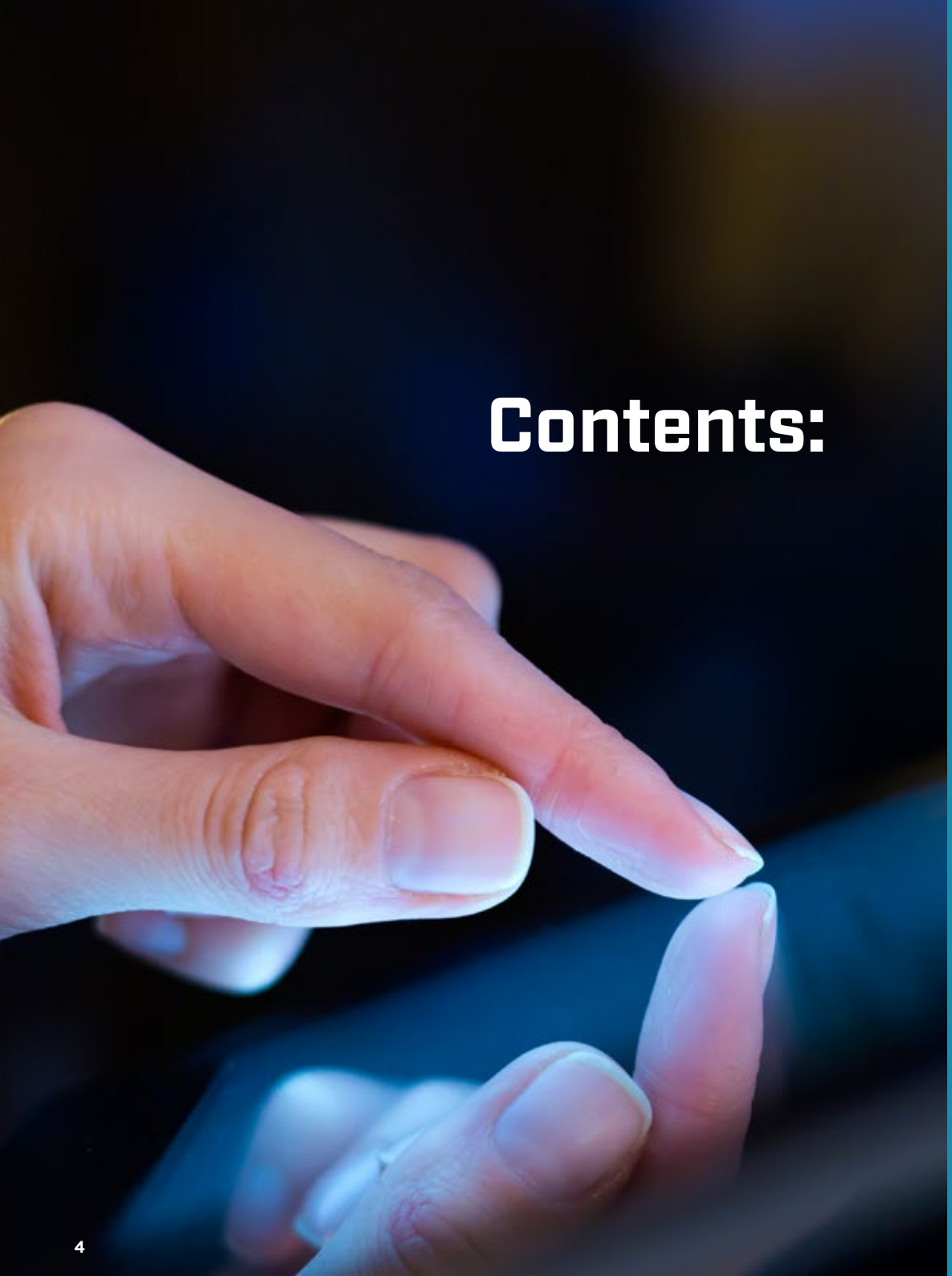
This document is primarily intended to be a tool for the program board when forming the program strategy, but the hope is that it also can help to unify the efforts of the participating companies and research groups. We also hope that other readers could find the foresight interesting and useful.

The research fields within WASP are moving forward rapidly. This document is, therefore, a snapshot and must soon be updated in order not to mostly be of historical interest.

It should also be mentioned that KAW after the production of the texts in this foresight have granted an additional billion SEK to WASP to expand the activities in AI. Although AI was already an important part of WASP, this added emphasis on the area is not reflected in this document.

The board expresses its gratitude to colleagues in industry and academia that have taken time out of their busy schedules to write valuable contributions to this foresight.

Mille Millnert, Chairman WASP

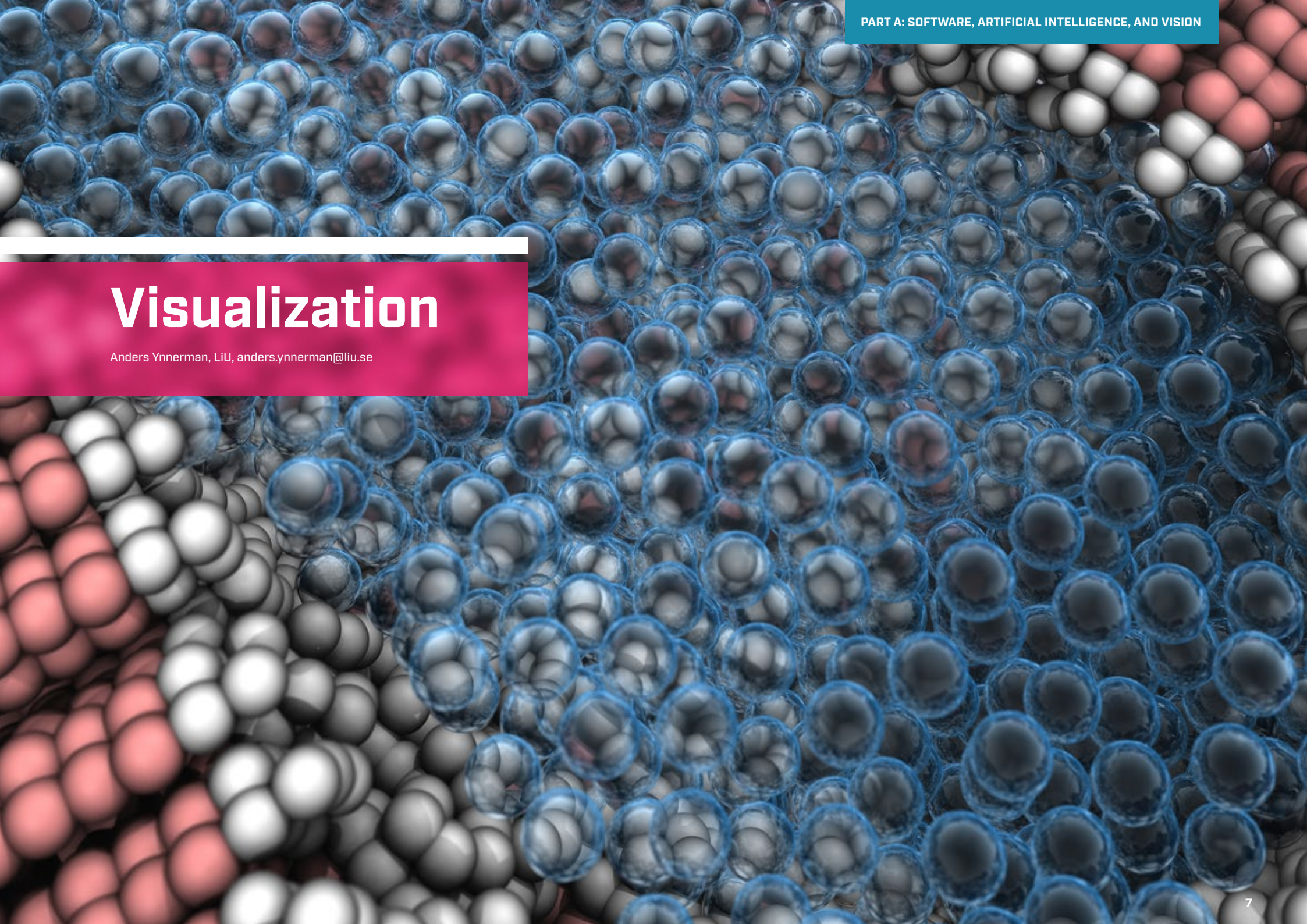


Contents:

Part B:
Smart systems, autonomous vehicles, and robotics 48

Part C:
Related trends and enablers 89

Part A:
Software, artificial intelligence, and vision



Visualization

Anders Ynnerman, LiU, anders.ynnerman@liu.se

OVERVIEW

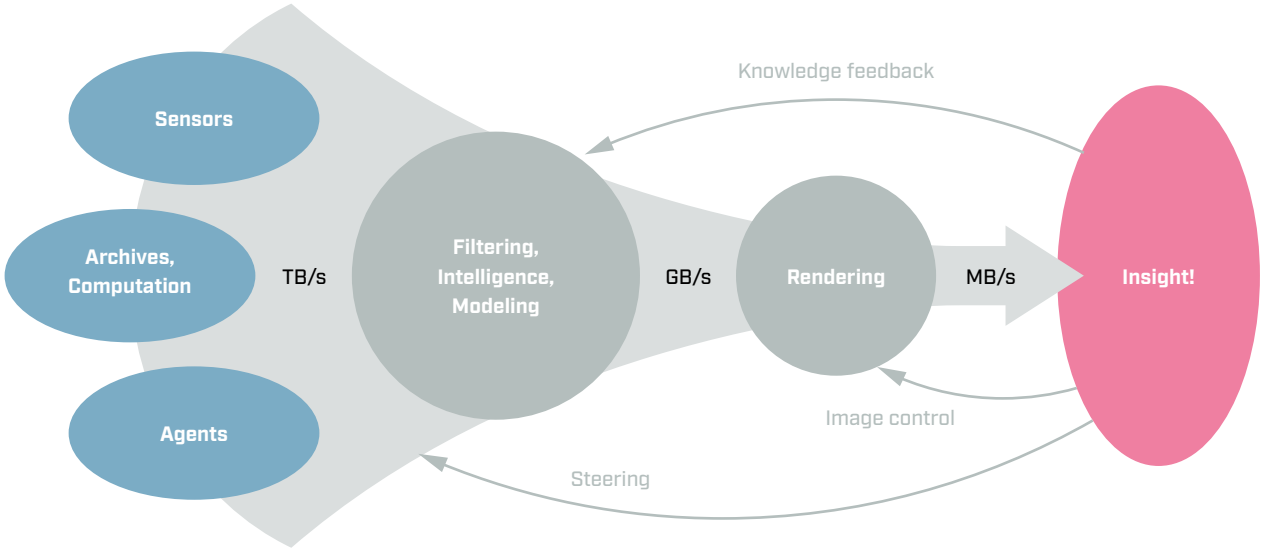
We are currently seeing a rapid increase in both the amounts and types of data available through sensors, from the internet, from databases and simulations and similar, which holds valuable information describing complex situations where active decision making by human experts is necessary.

Increasingly sophisticated autonomous systems are now collecting, analyzing and presenting data in aggregated forms. This precise information harvested through data analysis consists of Tb/s of streaming data that needs to be reduced to Mb/s for human interaction and decision making – a six-orders-of-magnitude decrease.

In order to solve this technical and societal challenge, it is of the highest importance to develop decision-support systems which adaptively reduce the cognitive load caused by large and rapid information flows while ensuring application-dependent mission-critical decision time-scales. This means that the time-scales for short-term decision converges with the time-scales for long-term decisions based on strategic data. The autonomy of agents or vehicles in a system (whether it is software or hardware) changes the user's/operator's interaction from real-time control (direct maneuvering) towards longer time-scales (directives).

Similarly, the wealth of extracted and aggregated multi-source data, from for instance sensors, internet or databases, will shift strategic long-time-scale decision making towards shorter, minute-update time-scales.

Driven by the rapid increase of computational power and development of supporting technologies for interaction new possibilities to use both stationary, handheld and wearable devices, for decision support, the full range of platforms for human interfaces to information and systems are appearing. This includes large-scale centralized arenas for collaborative decisions in operational settings to distributed systems with multiple users at different locations, and of course any interconnected mix of these environments in networked setups.



A visual analysis pipeline with user feedback loops. The key challenge is to enable instantiation of this pipeline for heterogeneous multiscale time-resolved data, where the data amounts are far beyond what brute-force methods can handle even with state-of-the-art technologies. In other words, the essential challenge is to achieve extensive data reduction in all stages of the analysis and visualization pipeline, while preserving the precision needed for the task at hand.

This rapid development calls for new decision-support approaches to effectively place humans in the decision loop. Display of rapid heterogenous information flows at multiple decision time-scales represent major challenges in visualization and interaction, breaking with traditional linear approaches.

There is thus an urgent need for the development of new integrated techniques for human interaction with multiple data sources such as multi-agent information-generating autonomous systems with variable level of autonomy and user control, interfaces to visual-data-mining sources and resources and on-line simulation capability and capacity.

This foresight is based on the following four assumptions:

- Humans will continue to play a governing role also in the era of an autonomy-rich environment.
- Data flows will have to be reduced to fit the capacity of human perception (human-style communication).
- Autonomous systems will be key components in reduction of human cognitive load to enable human-time-scale decision making.
- Increasingly effective and engaging advanced interfaces will be needed to support high-level human-style communication.

As a starting point for the description of future decision-support technologies we use three future scenarios:

Scenario 1: Centralized decision support – decision arena

This environment contains novel technologies that provide the human interfaces to the decision pipeline showed in the figure above.

A common scenario is a screenscape environment (a landscape with multiple display and interaction devices). The rapid development of display technologies will dramatically change the way displays are used for presentation of visual metaphors representing both abstract and spatial information and overlaid combinations thereof. Displays of the future will be automultiscopic, which means that they will support 3D stereos without glasses, have high dynamic range and retina resolution. There will also be transparent displays. All displays will also be multi-touch enabled.

The environment will be multimodal to enable multiple information channels to the human operator. Sound and speech synthesis will be integrated and tangible interfaces will also provide peripheral och contextual information. Human feedback can be based on a combination of speech, gestures, and haptic interfaces as well as traditional interfaces.

A more long-term development could be based on brain-computer interfaces providing a direct link between the system and the human operator. A challenge in this scenario is to balance the capability of the human perception with information flows and interaction possibilities.

Important aspects are also to provide visual representations that promote trust and human-style communication such as humanoid avatars that can serve as guides and assistants in the complex information flow.

Scenario 2: Personalized and contextualized decision support in the wild/field

At the same time as the centralized decision-support environment will reach unprecedented levels of sophistication and maturity, advanced handheld devices and VR/AR will create possibilities for personalized and contextualized decision support everywhere. Wearable devices will take the full step to become ubiquitous and pervasive and will enable control of communication flows and views of data and steer remote operations.

This means that visualization of data becomes seamless as users can access centralized data on-the-go through wearable technologies. In conjunction with increasingly accurate mapping, navigation- and tracking-augmented situational awareness will become a natural part of the future personalized and wearable decision-support system – the digital cognitive companion –, which will constitute a new level of human-technology relationship best described as a "sixth sense" aiding humans in complex situations.

Scenario 3: Collaborative decision support

A third scenario that deserves to be mentioned is the interlinking of different flavors of decision support involving human and cyber-physical agents in the field and operational centers for analysis and operational steering on one or several locations.

The challenge here lies in building interfaces that enable consistent and shared situational awareness and communication of strategies across the heterogenous environments. All data streams will not be available on all sites and in all contexts, but the high-level current state of the system, history and future evolution should be generally available .

A typical use case would be the emergency-response efforts before, during and after a terror attack such as recent incidents at the time of writing. In future situations, humans and autonomous systems, physical and virtual, will operate in a mixed initiative mode. Human interfaces to the information will play a crucial role in the success of these operations.

INTERNATIONAL MATURITY

Visual decision support has been identified as one of the key underpinning technologies supporting the ongoing information and intelligence revolution. Some aspects of the context have reached high level of maturity as it is based on a long progression of research and work practices, whereas other aspects are in their infancy. This is especially true for paradigms based on big data and interaction with intelligent and autonomous systems.

Visualization is a maturing field; visual representation approaches to effectively analyze a wide variety of data have been developed. Most of these methods are, however, surprisingly dealing with static data. Visualization of streaming data can in some cases take these methods as a starting point, but in general, new approaches are needed.

It is also noted that, as the field advances, visual representations are often tailored to specific applications and there is a lack of ontologies and classification of methods and their applicability. As a consequence of this, interaction schemes needed for feedback loops have not been not been given sufficient attention.

When it comes to the underpinning technologies, display technology is one of the fastest-developing technology fields at the moment and the speed at which new technology is brought to the market is unprecedented. This affects all of the decision-support scenarios using the range of display systems from large-scale displays and projections sytems to head-mounted virtual- and augmented-reality systems.

The topic also touches upon new and novel ways of communicating information using human metaphors such as physical and virtual avatars representing data, agents and intelligence in the decision support pipeline. This is an emerging field with a low level of maturity.

Some international players to be mentioned:
- SCI Institute, University of Utah;
- Technical University of Vienna;
- University of Stuttgart;
- Institute for Creative Technologies, UCLA.

CHALLENGES

The challenges involved in enabling the scenarios described above span across several disciplines. In this

foresight we focus on the human interfaces and assume that the interesting and difficult challenges of handling data and reducing it to fit representation in terms of visual metaphors and interaction schemes has taken place at earlier stages in the pipeline. It is, however, recognized that the challenges can not, and should not, be dealt with in isolation and thus multidisciplinary efforts are needed.

Supporting human-level interaction – cognitive companions

In many applications, remote autonomous systems will be embodied as avatars. In training, learning or personalized health care, the avatar could be manifested as a photo-real digital human. In other applications, such as big-data analysis or UAV-fleet management, the avatar, or cognitive companion, represents visualizations of multi-source aggregated data, or even advanced instrumentation in a driver environment.

New software architectures that can form dynamic application-defined networks with transparent use of underlying highly heterogeneous and changing physical networks are needed. These architectures further need to support high-level context-aware composition of autonomous services, evolvability of the software, and flexible security, supporting secure communication in the presence of dynamically changing networks and components.

Interaction with collaborative multi–agent autonomous systems at variable levels of autonomy will require formal specifications and techniques for symbiotic human–robot interaction processes in terms of delegation, contracts, speech acts, and other protocols as well as formal models for shared tasks used in associated with joint planning and decision making.

New visual representations for streaming heterogenous data

The range and heterogeneity of data will require new autonomous methods for choosing visual metaphors based on ontologies of visualization methods, and new data-dependent, adaptive, and suggestive visualization protocols.

As mentioned above, there is a lack of visualization targeted towards representing dynamic data. A challenge is to develop representations that effectively make use of the strengths of human perception in 4D and are tailored to the spatio-temporal resolution of human perception.

An interesting notion that can be exploited is the strong spatio-temporal correlation most data sets exhibit. This is an inherent feature as the characteristic time scale at which interesting events occur and are correlated through to the speed of information propagation. Thus regions of interest and level of detail in space and time are strongly correlated. This has so far not been extensively used in visualization. In summary, the challenge is to develop data-dependent, adaptive, and autonomous visualization, which, given the range and heterogeneity of data, selects visual metaphors to be used in visualizations based on ontologies of visualization methods.

Effective immersive data visualization

There has been renewed interest in the use of immersive technologies for data exploration. This has been spawned by the renewed interest in VR/AR and the improved possibilities that modern GPUs and improved display and tracking offers.

There are several ongoing studies on the use of head-mounted VR for data exploration, but the results seem to vary and are very application-dependant. There is also a renewed interest in collaborative visualization in immersive environments with large-scale displays; even dome-theaters and planetariums have been pitched as the best environment for collaborative visualization.

A challenge in all these environments is to find visual representations that are tailored to the needs of the application at hand and that make use of the immersive dimension to add clear value. For data analysis and decision support, immersion needs to be stable, controlled and reproducible, which is in sharp contrast to most current VR applications targeting gaming scenarios. There is thus a clear need to develop perceptual metrics for assessment of visualization and interaction quality in immersive environments.

TYPICAL INHERENT TECHNOLOGIES

Human interfaces to data and interaction with systems is a truly multidisciplinary field and there are several underpinning technologies that need to be addressed. A short list of some of the key areas for the described contexts are given here.

- Display technology: as described above, displays are evolving rapidly opening up new possibilities for human interfaces involving dynamic holographic representations.
- Tracking technology: in all scenarios, tracking of users, ranging from eye tracking to gestures, must be deployed.
- Novel tactile interfaces and other multisensory technology: this fills the need for more information channels and intuitive interaction.
- Speech recognition and synthesis: this is a necessary technology for human-style communication.
- Visualization of streaming data: research on new approaches is one of the most central topics for this foresight.
- AI and machine learning: even for the interfaces to humans, the high level of communication will have to be driven by intelligent systems and advances in AI; learning is required.

At the back end of the visualization pipeline described above are a multitude of technologies and research domains. The description of these is, however, beyond the scope of this foresight.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

As with most disciplines, developing tools for applications breakthroughs are often defined in terms of milestones of the application. Below is a list of potential application and technology breakthroughs:

- light-field GPU technology;

- glasses-free 3D displays;
- eyes-free control using multi-channel haptic on-body wearable tools;
- real-time synthesis of photo-realistic humanoid avatars;
- human-level interaction with multiple streaming data using virtual avatars expressing knowledge and analysis of the streams;
- learning systems for visualization design.

POTENTIAL FOR SWEDISH POSITIONING

Sweden holds a strong position in visualization in many application domains. There is now an opportunity to deploy this advantage to support the next generation of human-in-the-loop systems for decision support. A multidisciplinary approach capitalizing on research groups in visualization, computer graphics, interaction design and underpinning tracking, display systems in close collaboration with data analysis and artificial intelligence would enable the implementation of a state-of-the-art testbed for future decision support.

POTENTIAL PLAYERS

The leading site in Sweden for Visualization is at the Norrköping Visualization Center C, sorting under Linköping University. The research groups at the center are covering many of the areas outlined in this foresight.

The HCI research at Chalmers should be mentioned as a potential site and the HPC-related visualization at KTH is also closely related to the topics mentioned.

High-quality basic computer-graphics research is also found in Lund and at Chalmers.

NEED FOR SUPPORTING ACTIVITIES

Apart from funding of the research efforts outlined above, a testbed in terms of a large-scale laboratory for visual decision-support experimentation would be needed. In such a setup, it is not only the hardware that is needed, but primarily a significant and long-term pool of research engineers implementing and maintaining the testbed.

Deep learning and computer vision for autonomous driving

Jacob Roll, Autoliv, jacob.roll@autoliv.com
Erik Rosén, Zenuity, erik.rosen@zenuity.com

OVERVIEW

Traffic accidents is a large source of causalities and injuries with 1.2 million fatalities each year. Active safety aims at preventing, avoiding and mitigating accidents.

In recent years, the development of active safety has been extended to research and development aimed at autonomous driving. The concept of autonomous driving has not only the potential of drastically decreasing the number of accidents, but also the potential of redefining the way we think about automotive transportation (driverless freight transportation, mobility as a service (MaaS), car sharing and similar). The implications from such concepts on people's lives are difficult to grasp.

The road-transportation system is built around human drivers that largely base their driving on visual inputs. This, in combination with the relatively low cost of vision systems and recent advancements within deep learning and computer vision, makes it likely that vision-based technologies will play a prominent role in autonomous driving. Radar-based technologies will also be needed, but rather as a complement. Finally, lidar-based technologies are definitely interesting and may play an important role, but they are still suffering from issues related to cost, quality and mass-production.

All in all, we recommend that WASP has its primary focus on vision-based technologies, complemented by sensor fusion.

INTERNATIONAL MATURITY

Autonomous vehicles are being tested in real-life traffic in different areas of the world. Major tech companies (Google, Uber, Nvidia among others), start-ups (Comma.ai, Drive.ai, Zenuity among others), as well as established car manufacturers and suppliers (Daimler, Volvo, Tesla, Mobileye, Autoliv among others) are developing new technology and considering business opportunities in an unforeseen pace. Further, we read about mergers, acquisitions and strategic partnerships almost every day related to autonomous driving. Very few of these autonomous players have extensive experience of real-life safety.

CHALLENGES

Some of the major challenges one will need to face in the development of autonomous driving systems are:

- Robustness – for full autonomy, the system needs to be robust and work in all environments, and all weather and traffic conditions. It will be virtually impossible to foresee all situations that the autonomous vehicle can face and needs to handle in reality.
- Functional safety – the system should be able to make safe maneuvers in case a system failure occurs, in order not to cause unnecessarily dangerous situations.
- Cost reduction – for instance, it will be desirable to reduce the reliance on expensive sensor technology.

TYPICAL INHERENT TECHNOLOGIES

Due to its relatively low cost and rich information content, the vision sensor will play an important role for autonomous vehicles and active safety. Two prominent areas of research will be of particular importance.

1 **Deep learning** has revolutionized the area of machine learning in recent years, reaching new levels of performance in diverse applications. Until recently, most work on deep networks have focused on classification performance and not so much on computational efficiency. A general challenge is therefore to get computationally and power-efficient networks with robust performance. Some specific technologies relevant for autonomous driving include:
   - a Semantic segmentation/instantiation, where each pixel in the image is assigned an object class (pedestrian, vehicle, road, sign, building, vegetation and similar).
   - b Recurrent networks, where temporal dynamics of the surroundings can be taken into account.
   - c Deep reinforcement learning, where deep networks are trained to plan the path and possibly control the vehicle.
   - d Adding physical knowledge to networks, to efficiently capture geometric and other physical constraints.

   - e **Geometry-based computer vision** can be used to make a 3D reconstruction of the vehicle surroundings. SLAM/structure from motion is a family of techniques for simultaneously estimating the vehicle movement and the 3D structure of the environment, which is useful, for instance, for general object detection, navigation, dynamic camera calibration and general scene understanding. Sensor fusion can be used together with geometry-based computer vision techniques to combine information from different sensor setups (including stereo vision).

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

Deep-learning technology is booming, but until recently, not so much attention has been on real-time performance and computational efficiency. Focusing on network architectures that give the highest performance per numerical operation is one way to attack this problem.

Another way would be to incorporate physical knowledge into the networks, and/or combine deep networks with geometry-based computer-vision techniques such as SLAM. By combining these techniques, the deep networks should be able to focus on the parts that are difficult to model in more computationally efficient ways.

The connection to hardware is also interesting. How should hardware and software designs interplay to attain the most power-efficient performance?

Deep learning is still only in the beginning of its development, and there is great potential in many lines of research, including for instance recurrent networks and reinforcement learning.

SLAM is also interesting in its own right, and combined with navigation through cloud-based maps. On one hand, one interesting issue is to find efficient, low-bandwidth features for localization against the map. On the other hand, improvements of the actual SLAM functionality would allow for a sparser map representation.

Potential for Swedish positioning

Traditionally, Sweden has a strong background in many research areas relevant to autonomous driving (for instance vehicle systems and dynamics, automatic control, computer vision, signal processing/sensor fusion, system identification, crashworthiness and occupant protection). However, in addition to these traditionally strong areas, we see a need of applying deep learning. Today, research within deep learning is dominated by a small number of university groups (mainly in Canada, USA and the UK) plus major tech companies like Google, Facebook and Microsoft. We encourage a strong Swedish initiative in this area as well. That would probably include recruiting some of the most promising international researchers to Sweden. We believe that this could be made a reality within WASP and by connecting such efforts to the Swedish automotive cluster.

Furthermore, deep learning and, more generally, machine learning is closely related to mathematical statistics and system identification. Like its relatives, it consists of general techniques applicable to solve many real problems. This is accentuated by the fact that quite little domain knowledge is needed to develop and train deep networks, since they learn useful features automatically by stochastic gradient descent. It is therefore likely that a strong effort related to deep learning for the automotive industry will lead to advancements in other areas where deep learning could be applied.

POTENTIAL PLAYERS

To position Sweden as a strong player in deep learning, we envision building up an eco-system for deep-learning research. This may include establishment of completely new research clusters by recruitment of excellent international researchers, as well as strengthening existing research teams in related areas.

A natural spot for a strong research cluster in vision-based deep learning with autonomous-driving applications would be in Gothenburg, with its proximity to companies such as Zenuity, Autoliv, Volvo Cars and Volvo Group, which all have strong activities in this field, and Chalmers University, with its Computer Vision, Signal Processing, Computer Science and Mechatronics groups.

While there are apparent advantages of focusing efforts on building one strong cluster with co-located resources, attracting top researchers, there are also several reasons for complementing this with a broad effort (maybe 4–5 universities):

- To utilize, maintain, and further strengthen existing research groups in relevant areas. Examples (in addition to Chalmers) include:
  - The Computer Vision, Statistics and Machine Learning and Automatic Control groups at Linköping University.
  - The Mathematical Imaging group at Lund University.
  - The Computer Vision and Machine Learning and Automatic Control groups at Royal Institute of Technology (KTH).
  - The Systems and Control and Visual Information and Interaction groups at Uppsala University.
- To provide a wider range of high-quality study programmes in the field, and enable a more extensive recruitment of students.
- To provide several career opportunities in Sweden, in order to keep and attract promising young (as well as more established) researchers in the field.
- To stimulate a fruitful communication and collaboration between several sites.

By combining an initial focus on building a strong, internationally competitive deep-learning research cluster, with a wider effort of strengthening deep-learning activities on several sites, we believe that a strong and sustainable eco-system of deep-learning research in Sweden could be realized.

Need for supporting activities

There are two main resources that are crucial to develop successful deep-learning-based autonomous-driving systems:

- Data recording and annotation infrastructure.
- Deep learning training infrastructure.

Data recording and annotation infrastructure.

Deep learning is dependent on large amounts of training data. To be able to attain verified robust performance, data needs to be collected from a large variation of environments and conditions (including, but not limited to, different times of day, seasons, weather conditions, countries, types of roads, and traffic situations). To obtain ground truth, support from additional sensors such as LIDAR could be needed. Also, at least in an initial phase, large resources need to be spent on manual labelling of images.

Depending on detail, this could be a very time-consuming task. As a reference, the Cityscapes dataset[1] provides 5,000 images of "semantic, instance-wise, dense-pixel annotations of 30 classes", such as road, pedestrians, different kinds of vehicles and buildings. This kind of annotation took on average 1.5 h per image. This dataset covers German cities in daytime and in good weather conditions. To get the desired variation of conditions, one would need a several factors larger dataset.

There are two topics of active research that aim at reducing the amount of manual work for recording and annotation of training data. Firstly, various degrees of automatic labelling can facilitate the annotation work. Secondly, the use of synthetic data may be a way of getting a large amount of training data where the annotations come for free. Note, however, that the final system should be validated on an adequate amount of real data in order to be reliable.

Deep learning training infrastructure.

Training a deep network is a computer-intensive task. Currently, deep networks are most efficiently trained on GPUs, and frameworks such as TensorFlow, Torch, Theano and Caffe have been developed to facilitate this.

In addition to these frameworks, an infrastructure needs to be built up to efficiently handle scheduling of multiple training requests and the large amounts of training data needed.

Footnotes and links:
1. www.cityscapes-dataset.com/citation

Visual perception and deep learning for autonomous vehicles

Michael Felsberg, LiU, michael.felsberg@liu.se
Kalle Åström, LU, kalle@maths.lth.se

DEFINITIONS

Autonomy: "Choice to make free of outside influence". Automatic means that a system will do exactly as programmed, it has no choice. Autonomous means that a system has a choice to make free of outside influence, i.e., an autonomous system has free will. Brian T Clough, "Metrics, Schmetrics! How The Heck Do You Determine A UAV's Autonomy Anyway".

UXV: Unmanned vehicles in any environment: air, ground, on water, in water.

AXV: Autonomous vehicles in any environment: air, ground, on water, in water.

WARA: WASP Autonomous Research Arena.

WARA-CAT: WARA Collaborative and Autonomous Transport.

WARA-PS: WARA Public Safety.

OVERVIEW

The area of autonomous vehicles is subdivided into different sub-areas, originating from the different environments:

air – unmanned aerial vehicles (UAVs), example: drones;

- **ground** – unmanned ground vehicles (UGVs), example: autonomous cars;
- **on water** – unmanned surface vehicle (USVs), example: boats;
- **under water** – unmanned underwater vehicles (UUVs), example: sub-marines.

The vehicles in these four sub-areas can be collectively titled UXVs. In the process of increasing autonomy in these unmanned vehicles, UXVs are successively turned from teleoperation (or co-operation in case of cars) into fully autonomous vehicles, which are collectively called AXVs. Computer vision has been identified as a central capability for those systems[1].

Computer vision aims at computational methods (for instance optimization, inference and machine learning) and software to extract information from images and image streams acquired by cameras. For autonomous vehicles, computer vision is the key technology for visual perception of the vehicle's environment, turning the vehicle's cameras into systems for harvesting relevant information.

The term "Internet of Cameras" denotes the analysis and generation of meta-data from cameras. Examples here are surveillance, smart cities, structure from motion, cooperative map-making and similar. Cameras can be used as sensors to detect object, recognize objects, localize objects and interactions. Examples are:

- cameras that automatically detect free and occupied parking spaces and sharing such data in open data initiatives, for the benefit of users, municipalities, parking owners;
- cameras that automatically detect traffic users and analyze their behavior so as to assess and improve on traffic safety;
- cameras in cars that use build maps of cities, roads and similar cooperatively;
- cameras that measure the flow of people in cities, stores;
- 3D modelling of objects for measuring.

Both contexts are relevant to both WARA-CAT and WARA-PS. A major further area not covered by WARAs, but highly relevant for Sweden, is precision farming, both for live stocks (for example fish farms) and crop farming, also including foresting.

INTERNATIONAL MATURITY

Whereas UXVs are current state-of-the-art, AXVs and collaborating fleets of those are still subject to research, with many unsolved problems, including many vision problems. Artificial vision systems are still subject to failures far beyond acceptable level, leading to serious incidences[2]. Adverse vision conditions occurring due to weather and underwater vision are topics that have just recently been started to look at by the scientific community[3]. A huge leap forward has been achieved by the development of powerful deep networks.

The internet of cameras context is partially mature, but still underdeveloped. This applies to both legal frameworks and technological issues. Whereas legal issues highly influence society and industry, practical problems of camera networks "in the wild" require novel scientific solutions to develop suitable techniques.

Assessing players in the context is difficult as major activity in companies and military/defense is completely opaque or watered-out by PR strategies. Most activity is currently observed on the subfield of self-driving cars with companies such as Google, Apple, Tesla, Daimler, Volvo, all of which using vision sensors. Other areas of unmanned or autonomous vehicles are much less mature and even realistic benchmarking is not available, compare for example KITTI, using real data[4], versus UAV-benchmarks, using synthetic data only[5]. The situation is even worse for the internet of cameras context, since companies such as Facebook and Google usually do not reveal their state of development other than through their products and national security institutions do not disclose their status at all.

Alliances with for example Google, Apple, Daimler, Volvo, NVidia, Saab and NFC exist. These alliances clearly show the relevance of Swedish computer vision research for major companies working on autonomous vehicles and public safety.

CHALLENGES

For the general area of autonomous vehicles, major challenges to overcome are[6]:

- how to reduce the level of supervision during learning;
- the reduction of required data for the learning;
- learning of layered dynamic processes;
- the verification of learning based systems.

When it comes to camera technology, there are some legal and technological challenges:

- privacy, regulations and legislation;
- asynchronous streams (unknown relative timing) including frame-drops;
- unknown localization and perspective, unknown camera internals (for example focal length);
- color distortions and blur;
- the vast amount of data to be transferred and processed.

TYPICAL INHERENT TECHNOLOGIES

In the area of visual perception for autonomous vehicles, a number of typical technologies and research areas can be identified and highlighted as being of particular interest for the future.

Machine learning

The area of machine learning[6] is currently largely dominated by the deep-learning approach. This approach has led to major progress, but several issues remain open:

- deep learning – weakly supervised, un-supervised, or by reinforcement learning;
- the injection of hard constraints, structural and geometric structure, regularization and invariance during learning as well as synthesizing and augmenting data;
- learning in recurrent networks and with layer-wise control;
- assessment of performance on system level, systematically varying, non-linear.

SLAM/SfM/visual odometry

One research area is the problem of Simultaneous Localization and Mapping (SLAM), Structure from Motion (SfM) and/or visual odometry. Within this area, one uses sensor data for example images, sound, radio, accelerometer, gyros, in order to simultaneously estimate parameters of the sensors, such as position, orientation, calibration parameters, but also parameters of the environment, such as 3D shape of objects, buildings, sound sources, wifi-base station positions and similar. These problems are non-linear parameter-estimation problems involving measurements error, outlier errors and missing data.

Object- and action-recognition, image understanding

This research area involves the question of developing algorithms for classifying and recognizing objects and actions to find semantic relations in a scene and finally to derive a complete understanding of the contents of images. This technology area can be considered as the semantic counterpart of SLAM/SfM/visual odometry, which is focusing on geometry. The area is currently developing fast with new machine-learning techniques such as deep learning and requires functioning detection and tracking methods – see below.

Object and action – detection and tracking

Image sensors deliver distribution of light intensities at various spectral wavelengths, but methods from SLAM/SfM/visual odometry and object- and action-recognition require spatial positions or trajectories of objects, landmarks and actions. Detection and tracking focusses on the localization of object hypotheses in the signals generated by cameras. The following issues will have to be addressed:

- systematic and application-relevant evaluation of detection and tracking;
- significant increase of robustness of methods using results from deep learning;
- self-assessment of methods, detection of failures and determination of confidences;
- moving from bounding boxes to pixel- or point-wise localization (segmentation);
- reduction of computational demand (time or energy).

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

Deep learning has initiated a breakthrough[7,8] in the field of computer vision. However, the process is far from being completed. Deep-learning algorithms have already been used in the 90s, but it was first in 2012 with the availability of the ImageNet dataset (14 million images, 1 million with annotation), GPUs (and NVidia CUDA language) and the use of suitable methods (ReLu-activation, max-pooling, drop-out) that deep-learning-based methods became state-of-the-art.

However, progress is still very steep and most other relevant areas suffer from the absence of ImageNet-sized datasets or energy consumption constrains that do not allow the use of GPUs (for instance battery-powered vehicles).

Also, very little is understood about why Deep learning works as well as it does. Most questions raised above are in one way or another related to improving the modelling of deep learning. First when methods for predicting the performance of Deep learning exist, the previously mentioned problems can be addressed in an optimal way. Presumably, the future improvement of deep-learning methodology will lead to a more significant improvement than the progress seen so far.

POTENTIAL FOR SWEDISH POSITIONING

Sweden has, more or less, missed the deep-learning revolution up to now. There are many reasons why this happened, starting from a strong disbelief in machine learning on the industrial side, lack of public research funding in the relevant areas, up to local protectionism of established research groups at Swedish universities and individual researchers failing to see that the new development is more than just another trend.

EXAGGERATED TRL

Regarding TRL, both researchers and industry typically overestimate the state of development. This happens partly due to general proposal-writing polemic, but also due to simplified statements from the PR divisions.

A major problem is that means for reliable assessment of state-of-the-art (benchmarks) are still under development. Most benchmarks used today have been started during the past decade (see KITTI and the UAV benchmark above, VOT[10] and LSVRC[11]) and progress is still very steep but a relatively low level.

For instance, expected localization accuracy measured in area overlap is improving by about 8%-points per year, but the current top score is at 33% (see VOT 2016). Thus, the general TRL of computer vision areas on data "in the wild" is no better than TRL2: the level of basic research and research to prove feasibility has just started.

Obviously, certain niches have achieved much higher TRLs, up to operation (TRL9), but this is typically achieved on very specific use-cases on highly regular data.

As argued above, the largest part of the revolution is still to come and thus, Sweden still has a chance to catch up with the development. Several researchers have started to work on deep learning some years ago; a community starts to form (for instance, the first symposium on deep learning took place in June 2017[9]), the funding situation improves (also due to WASP) and industry has directed most research to deep learning.

The possibly largest threat to stop the development is local protectionism, making too large a portion of funding end up in competing areas and leading to a funding level for deep learning below a critical threshold.

POTENTIAL PLAYERS

Pointing out specific potential players at this point might lead to directly preferring certain companies, universities/research groups/individual researchers in too early a stage of development. WASP has, already due to its size, potential major impact here, up to the level that the success or failure of deep learning in Sweden might depend on WASP. See suggestions below.

NEED FOR SUPPORTING ACTIVITIES

In order to identify central players in this early phase and to answer the question above, it might be suitable for WASP toset up a small group as a task group on deep learning in Sweden. This task group should then:

- observe activities in the field;
- make polls among involved players;
- organize workshops.

As objective of these activities, suitable players are to be identified, supported by WASP and connected to the relevant arenas, for example within smart cities. The last objective then links back to the context in terms of the WARA-X use-cases, to societal challenges and finally to the benefit of Sweden.

Footnotes and links:

1. www.vision-systems.com/articles/print/volume-22/issue-2/features/industry-solutions-autonomous-vehicle-driven-by-vision.html
2. arstechnica.com/cars/2016/09/mobileye-and-tesla-spat-heats-up-as-both-companies-trade-jabs
3. www.dagm2011.org/adverse-vision-conditions-challenge.html, http://marinerobotics.eu
4. www.cvlibs.net/datasets/kitti
5. ivul.kaust.edu.sa/Pages/pub-benchmark-simulator-uav.aspx
6. dx.doi.org/10.4230/DagRep.5.11.36 (section 5.3)
7. The Deep learning Revolution: youtu.be/DyOhJWItsyE
8. The Rise of Artificial Intelligence through Deep Learning | Yoshua Bengio | TEDxMontreal: youtu.be/uawLjkSI7Mo
9. ssdl2017.se
10. www.votchallenge.net
11. www.image-net.org/challenges/LSVRC/2014

Machine intelligence

Elena Fersman, Ericsson Research, elena.fersman@ericsson.com
Rickard Cöster, Ericsson Research, rickard.coster@ericsson.com

OVERVIEW

Ericsson's definition of machine intelligence (MI) is combination of artificial intelligence (AI) and machine learning (ML). Academic research in the areas of AI and ML stretches over 50 years back. It has been successfully applied in many contexts ranging from health to social networking.

The field is very broad and the applications mostly use simple algorithms, that work especially well on unstructured data such as raw text (extensively used by companies such as Google and IBM). Expert systems assisting humans in decision-making use graph-traversal techniques[1]. Recently, MI has been used for recommendations, by building up a knowledge graph in any field (such as commerce and social networking) and constantly improving it through ML.

The ultimate benefits of the field of MI is automation as well as amplification of human tasks where the end result of combined human-machine effort becomes better than if it was performed by each of them separately.

INTERNATIONAL MATURITY

MI is being applied in all fields today given its clear benefits. The leading players are big software companies like Google and IBM whose business models depend on the smartness of their algorithms. However, only a small part of MI algorithms are proven to be applicable commercially. Many theoretic methods need further studies in scalability and efficiency to be applicable in practice.

CHALLENGES

Digitization[2], meaning the process of changing from analog to digital form, is a prerequisite for any application of MI, and many fields are not yet fully digitized. Another challenge is related to the security, safety and privacy of data and knowledge; non-willingness to share data and meta-data often slows down research advancements. Heterogeneity of data and knowledge is also a challenge when applying MI methods.

Major progress in recent years has been made in the areas of deep learning (DL) and reinforcement learning (RL)[3].

DL is performing very well for certain tasks, in fact especially such tasks where there have been available large open data sets for researchers, such as ImageNet[4], since researchers have then been able to construct new DL architectures suited for those problems. The progress has been made possible by increased computational resources and algorithmic advances, and relies heavily on the availability of data. To further develop the applicability of deep learning, universities and research organizations need not only theoretical research but also access to data. Today, data is first and foremost owned by large corporations that use this data for their competitive edge, both for developing new differentiated solutions and for pushing boundaries of academic research. Academia do not generally have access to this data, and therefore the larger corporations have an upper hand in MI research, which becomes a challenge for academia.

RL is becoming practically feasible for automated control but still requires extensive engineering effort for each case. Current work on the use of RL for automatic control relies on extensive engineering, for example deciding how an RL algorithm is allowed to do exploration in the system. Moreover, RL typically requires many iterations which may not be feasible in certain systems. A relevant avenue of research is to try to reduce the amount of learning iterations by different means, and to research how to combine simulators and real systems. It is also important to consider how human decisions or input can be used in guiding learning, as this could potentially broaden the use of RL for automation.

TYPICAL INHERENT TECHNOLOGIES

While many algorithms developed throughout the years are now possible to execute thanks to enhancements in supporting technologies such as distributed execution and major increases in memory capacity, computational power and advances in computer architectures, there is a need in heuristics and optimizations allowing not only brute-force execution but also intelligent real-time algorithms for walking through large amounts of interconnected knowledge and data. In other words, the system is supposed to be able to make decisions during its run-time, with a given certainty, without going through all information available.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

Constant improvements in computational capacity, architectures for data processing, and distributed cloud are building blocks allowing for efficient MI processing. Recent developments in blockchain technologies allow for secure distribution of data to be processed.

Successful application of MI algorithms in all fields will lead to high degree of automation, as well as machine-assisted innovations applicable to Internet of Things (IoT), 5G and service-delivery domains.

POTENTIAL FOR SWEDISH POSITIONING

The Swedish research community has historically been strong in the areas of formal methods for modelling and analyzing complex systems. This research has now developed into the field of cyber-physical systems that focuses on the interactions between the physical and digital worlds, taking networking and system-of-systems aspects into account. Machine intelligence is a key area that supports research in cyber-physical systems, allowing the system to learn and act in an intelligent way. Research in real-time decision making for cyber-physical systems taking safety, security, privacy and trust into account is therefore of importance and high potential for Swedish positioning.

There is extensive research on DL and RL in the world. Sweden could position itself by advancing DL and RL in

areas where Swedish industry is strong, for example in telecommunications and manufacturing. Advancing RL and DL for automated control of near-real-time systems is relevant to both telecommunications and manufacturing. For many problems, it is also desired that the ML model outputs prediction confidence. There is some work in this area in Sweden, for instance using conformal prediction and Bayesian techniques, but more focus could make for instance DL even more applicable to automation and optimization problems.

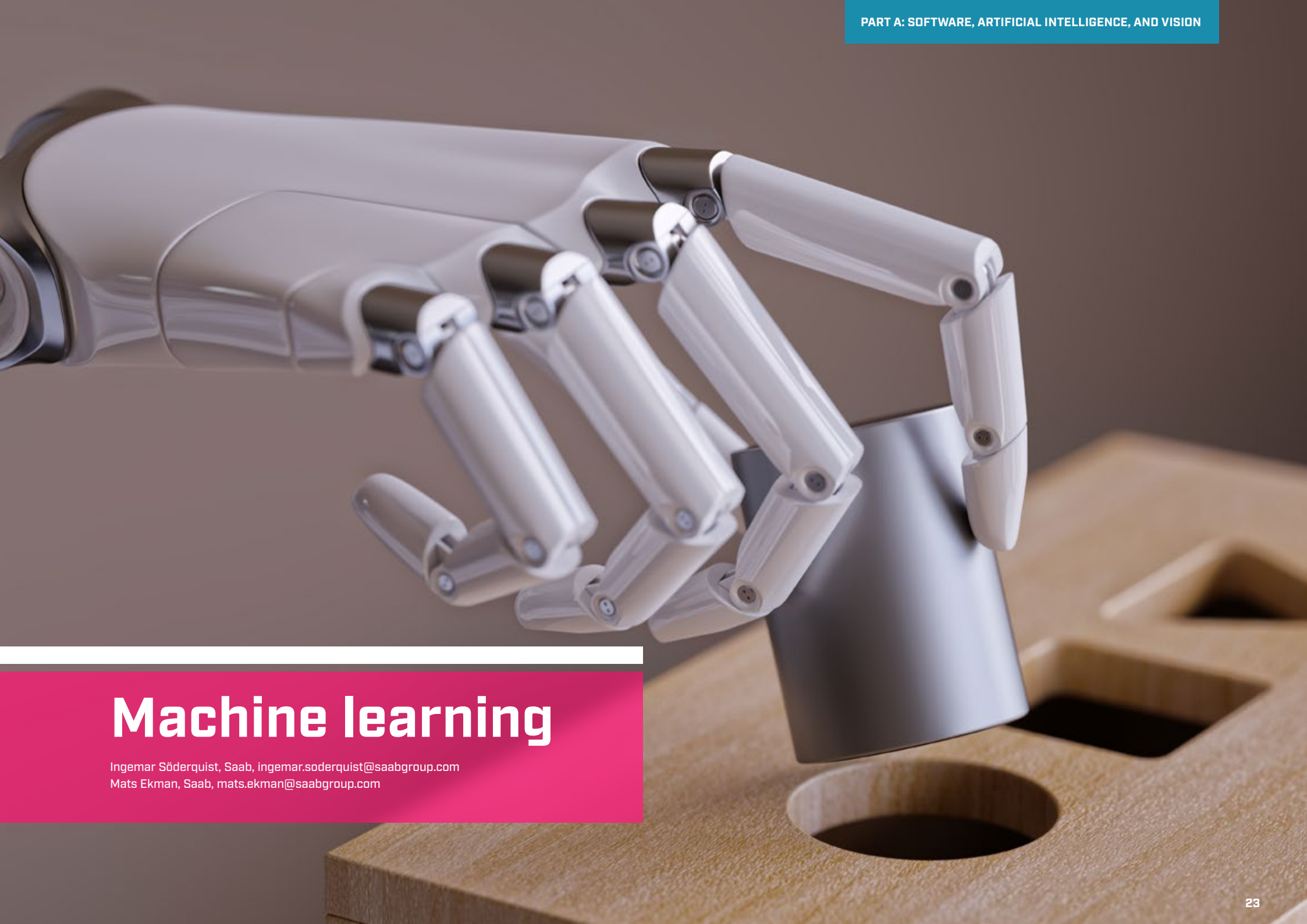
POTENTIAL PLAYERS

The field of MI is applicable to all WASP partners. In order to boost Swedish research and innovation leadership in the area of MI, we recommend creating an alliance between Swedish industrial and academic partners with a purpose of creating a common data and knowledge lake and enabling higher degree of interaction and innovation using this cross-domain data.

NEED FOR SUPPORTING ACTIVITIES

It is important to have increased collaborations between industry and academia, as well as efficient dissemination of results and information sharing. Data and knowledge lakes as well as common knowledge repositories stretching across academic and industrial partners can serve as a technical enabler for such collaborations.

Footnotes and links:
1. en.wikipedia.org/wiki/Graph_traversal
2. fersman.blogspot.se/2016/06/digitalization-digitization-and.html
3. en.wikipedia.org/wiki/Reinforcement_learning
4. en.wikipedia.org/wiki/ImageNet



Machine learning

Ingemar Söderquist, Saab, ingemar.soderquist@saabgroup.com
Mats Ekman, Saab, mats.ekman@saabgroup.com

DEFINITIONS

AI: Artificial intelligence, intelligence displayed by machines.

MI: Machine intelligence, see AI.

ML: Machine learning, the ability of computers to learn without being explicitly programmed.

OVERVIEW

Machine learning (ML) is an important building block for artificial intelligence (AI), or machine intelligence (MI). AI is defined as intelligence displayed by machines, in contrast with natural intelligence displayed by humans or other animals[1]. The field of AI research was founded as academic discipline in 1956. In computer science, AI research is defined as the study of "intelligent agents" (for instance autonomous aircraft), any system that perceives its environment and takes actions that maximize its chance of success at some goal. Traditional AI research includes reasoning, knowledge, planning, learning, natural language processing, perception and the ability to move and manipulate objects.

ML is defined as giving "computers the ability to learn without being explicitly programmed" attributed to Arthur Samuel in 1959[2]. Two major categories of ML are supervised learning in which the algorithms are "trained" by "training" data set of inputs and outputs, and unsupervised learning in which a data set is not a-priori structured and the algorithm finds structures and hidden patterns in the data.

Supervised learning can be applied to find the best algorithmic fit to a given data set, thereby enabling prediction of the output from a new data point.

Unsupervised learning can be applied as a form of data mining in unstructured data sets to reveal otherwise unknown structures and relationships in the data set.

INTERNATIONAL MATURITY

ML is nowadays applied in numerous ways in various industries and is receiving significant interest and investment in the area of autonomous self-driving automobiles. ML is also getting significant interest from defence laboratories and industries.

ML can be utilized in many different contexts such as route planning and replanning, sensor fusion, decision support, target detection, detect-and-avoid and similar. It can also be used to analyse patterns in big data, for instance to detect cyber-security attacks that affect typical communication patterns or patterns in health-monitoring data to find trends.

CHALLENGES

Complex aircraft systems are today developed using rigorous processes for system-, software- and hardware-development assurance such as SAE ARP 4754 Rev A[3], DO-178C[4], and DO-254[5]. These assurance processes assume deterministic behaviour and traceability from requirements down to actual design (for instance source code). With ML training, there is no traceability to the detail ML software implementation (for example neural-network weights) and lack of deterministic behaviour, consequently it is not possible to comply with all of the DO-178C or DO-254 objectives. Therefore, new verification methods are required for ML algorithms.

FAA/EASA parts 23, 25, 27, and 29[6] all state "the equipment, systems, and installations must be designed and installed to ensure they perform their intended functions under all foreseeable operating conditions". Accordingly, process for automotive domain is ISO 26262[7]. If ML is used for example to increase the level of flight-control autonomy, how can an applicant demonstrate that the flight control performs its intended function under all foreseeable operating conditions? The operating conditions are infinite and may need to include rain, snow, dust, viewing angles, sensor imperfections, lighting changes, system failures, and similar.

Automotive industry stands before similar technology challenges as avionics industry. ML techniques can give huge benefits especially in the area of self-driving cars. However, due to less developed regulations compared to aerospace, the challenges and research needs are even larger.

ML is considered to be the cutting-edge in software technology and is associated with intellectual properties (IP) with potential high market value. Therefore, it is very unlikely that ML suppliers will provide detailed information of the implementation to the system integrator to support safety assessment of the product including the IP.

Furthermore, the suitability to apply ML technology developed for one domain into a new domain is not easily analysed by one part only (ML supplier or system integrator).

TYPICAL INHERENT TECHNOLOGIES

The meaning of ML as an acronym needs to be clearly described. Today the term is often used as a buzzword in marketing of products or companies without any deeper description. The use can be for example to replace or support steps in the development process (see ARP above) or being a feature in the product itself.

Certification aspects are important to understand. In aeronautics, and probably in car industry, the use of ML during development needs to produce certifiable designs, which means that some proof of rigor development process needs to be justified by the developer. Today, use of ML cannot be certified due to the lack of acceptable means of compliance. This needs to be defined by the certification authorities.

Furthermore, when ML is part of the product, the product will be able to learn something not well known during the development of the product, and is therefore not verifiable as part of the product design base. This is conceptually quite different compared to current system-design processes used in avionics. ML can be considered an intended functionality in the product to be certified. However, the intended function needs to be deterministic during the lifetime of the product.

Further research is needed to determine – and understand – the level of training and amount of correct data needed to guarantee correctness in the software function to the same level as traditional algorithmic development.

Further maturation is needed of current methodology and optimisation of resources needed for the learning stage, both supervised and unsupervised, to wider use of ML technology in mobile devices and other smaller platforms or vehicles, compared to commonly used cloud technology based on processing power in large centralized computer clusters.

Research is needed in new combined software and hardware architectures that efficiently process data during the "inference workload" during learning.

There are several initiatives for new architectures spanning from extreme solutions to products for broader use, such as:

- tensor-flow processors, open-source software libraries and dedicated hardware[8];
- Intel AI portfolio and platforms[9].

Common is that the new architectures have more memory easily accessible from the CPU and efficient accelerators that offload the CPU/ALU. Also, an intuitive and easy-to-use tool chain, supporting the new architecture of combined soft- and hardware, is needed.

Several industries that successfully apply ML and AI settle with testing when ensuring confidence in their products. When applying recommendations from the avionics community, additional analytical verification activities are performed. In some cases, such verification activities may be automated by tools. In other cases, they must be performed manually by a domain expert. The avionics community needs more information on how to reason about correctness in software functionality.

Below are areas we believe must be addressed before vehicles based on ML technology are globally accepted. A more systematic analysis of state-of-art resulting in a focus on technical research areas and development goals needs be carried out.

- During development, using ML together with traditional methods.
- During development, finding verification methods not based on requirement breakdown.
- Methods to verify autonomous systems that include ML.
- Methods to determine deterministic behaviour of learning phase in target system.
- Methods to determine deterministic behaviour of cooperative systems (SoS).
- Configuration management of ML behaviour.
- Integrated maintenance based on ML.
- Maintenance based on big data and ML (trend analyse).

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

We need to identify and understand many, not only technical, aspects of how the use of ML differs from software of today. Software based on ML contains inherent uncertainties and behaves different compared to standard algorithmic software. For example, a car with automated parking uses rear camera for obstacle detection in the proximity; ML image detection would result in probability results "I am 90% sure that it is a free parking slot between two cars", and also the learning can be based on wrong data leading to unpredictable results in future situations.

POTENTIAL FOR SWEDISH POSITIONING

All areas above give knowledge of ML contribution as technical or as customer view of benefit in the product (for example car or aeroplane). All Swedish companies need to have this knowledge to develop, produce and deliver competitive products to the market.

Sweden is a relative small country; therefore research areas that focus on society's acceptance for this technology will benefit from collaboration with other countries in Europe.

POTENTIAL PLAYERS

Best possibilities have players with complete system knowledge, including the use of AI, and where all parts of the product are developed in Sweden. Examples are the traditional aeronautic, car and truck industries as well as the new and expanding game industry.

Swedish aerospace industry has potential for utilizing ML in both unmanned and manned avionics systems. The players currently involved are likely to be able to play a role in the future, including for instance the many start-ups gathered in the Swedish Aerospace Cluster.

Autonomous driving of vehicles can benefit from the structured approach to safety that aviation has taken. Swedish automotive industry including suppliers will be having opportunities here. Many industries increasing automation for vehicles or industrial systems with demanding requirements on safety and reliability have similar opportunities. If this proves to be true, then we could see possible contributions of various importance to a large portion of Swedish system-building industries.

NEED FOR SUPPORTING ACTIVITIES

The most important supporting activity would be to establish trust and acceptance of this unpredictable behaviour of ML with the end user of the product, and with society

in broad sense. For example, today the car driver is always responsible. In the future, with increased use of AI and ML, it is not clear whom to blame if, instead of parking correctly, by mistake the car injures a child in baby carriage; will the society argue that the responsibility is with the algorithm developer, the product developer or the user of the product?

International harmonisation is needed. In the long-term perspective, this might affect the entire insurance business. An indicator for fulfilling this achievement is probably that a common understanding has been established in society, academia, and industry.

Footnotes and links:

1. en.wikipedia.org/wiki/Artificial_intelligence
2. en.wikipedia.org/wiki/Machine_learning
3. ARP4754A, S. A. E. “Guidelines for development of civil aircraft and systems.”, December 2010.
4. RTCA/DO-178C, Software Considerations in Airborne Systems and Equipment Certification. RTCA, 2011.
5. RTCA/DO-254, Design Assurance Guidance for Airborne Electronic Hardware, RTCA, 2000.
6. www.easa.europa.eu/regulations
7. ISO, “26262: Road vehicles-Functional safety.” International Standard ISO/FDIS 26262, 2011.
8. www.extremetech.com/computing/247199-googles-dedicated-tensorflow-processor-tpu-makes-hash-intel-nvidia-inference-workloads
9. www.intel.com/content/www/us/en/analytics/artificial-intelligence/overview.html

Autonomous capabilities of software systems

Martin Monperrus, KTH, martin.monperrus@csc.kth.se
Benoît Baudry, KTH, baudry@kth.se

OVERVIEW

This text focuses on two kinds of autonomous capabilities of software systems: self-healing and self-diversifying. These capabilities are essential for achieving true reliability and availability in ultra-open environments.

INTERNATIONAL MATURITY

Code-level approaches for autonomous software systems is a completely new research field, which has emerged along the growing scale of software systems. It is at the forefront of research endeavors that emerge at the boundary between software-engineering, programming-language and operating-systems research. It is currently investigated by a handful of pioneering scientists in North America and Europe.

WORLD-CLASS LEADING PLAYERS

Martin Rinard (MIT, USA); Emery Berger (UMass, USA),; Stephanie Forrest (Arizona State, USA); Antonio Carzaniga (Lugano, Switzerland); Cristian Cadar (Imperial College, UK); Arie van Deursen (TU Delft, NL).

CHALLENGES

The challenges of self-healing and self-diversifying software are: scale, implicitness and heterogeneity.

- By **scale**, we mean that any high-value software system is composed of millions of lines of code. Automated reasoning on code at that scale is extremely hard; it would takes centuries to execute automated reasoning on software of the size of the Mozilla Firefox browser or the Linux Kernel.
- **Implicitness** concerns the available specification: the knowledge about how software should behave is often partial and buried into a variety of documents that cannot be automatically processed by other programs with the state-of-the-art of natural language processing and code analysis together.
- Finally, due to a high level of reuse, all major software systems use a set of **heterogeneous** software technology (for instance different programming languages). This makes it really hard to apply a powerful software-analysis technique that works on a single technology over multiple heterogeneous software stacks.

TYPICAL INHERENT TECHNOLOGIES

This research is built on key technologies for software analysis and transformation. These technnologies automatically reason about and manipulate program code.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

We foresee two breakthroughs, which can emerge from current research efforts. These are two extremely ambitious objectives for scientific research and tool development, which have the potential to fundamentally change the way software systems are built and can provide highly dependable services to our society:

- **Moving-target defenses.** There is currently a major asymmetry between attackers and defenders in software systems: defenders build systems that are complex, yet, rarely change; which gives a major edge to attackers who can spend time finding one vulnerability that they can exploit on millions of copies of a single program. Moving-target defenses shall integrate advances in code transformation, runtime deployment and automatic diversification to let software systems constantly change.
- **Antifragile distributed systems.** Internet-scale autonomous software exhibit certain behavior that cannot be observed in the lab and only happen in production. Consequently, it is essential to rethink the software verification stack. Antifragility shall exploit research results in the areas of runtime fault injection, online monitoring and online patch generation in order to set new foundations for the way software companies apprehend correctness and dependability.

POTENTIAL FOR SWEDISH POSITIONING

The next unicorns in software technology shall be about advanced technology for autonomous software, such as autonomous software diversification and chaos engineering. The very first startup companies in these fields have been founded in 2017 in the San Francisco Bay Area, USA (Polyverse1 and Gremlin2). It is still time for Sweden to invest in companies that develop game-changing software technology.

POTENTIAL PLAYERS

Key players to develop technology for autonomous capabilities of software systems are present in Sweden:

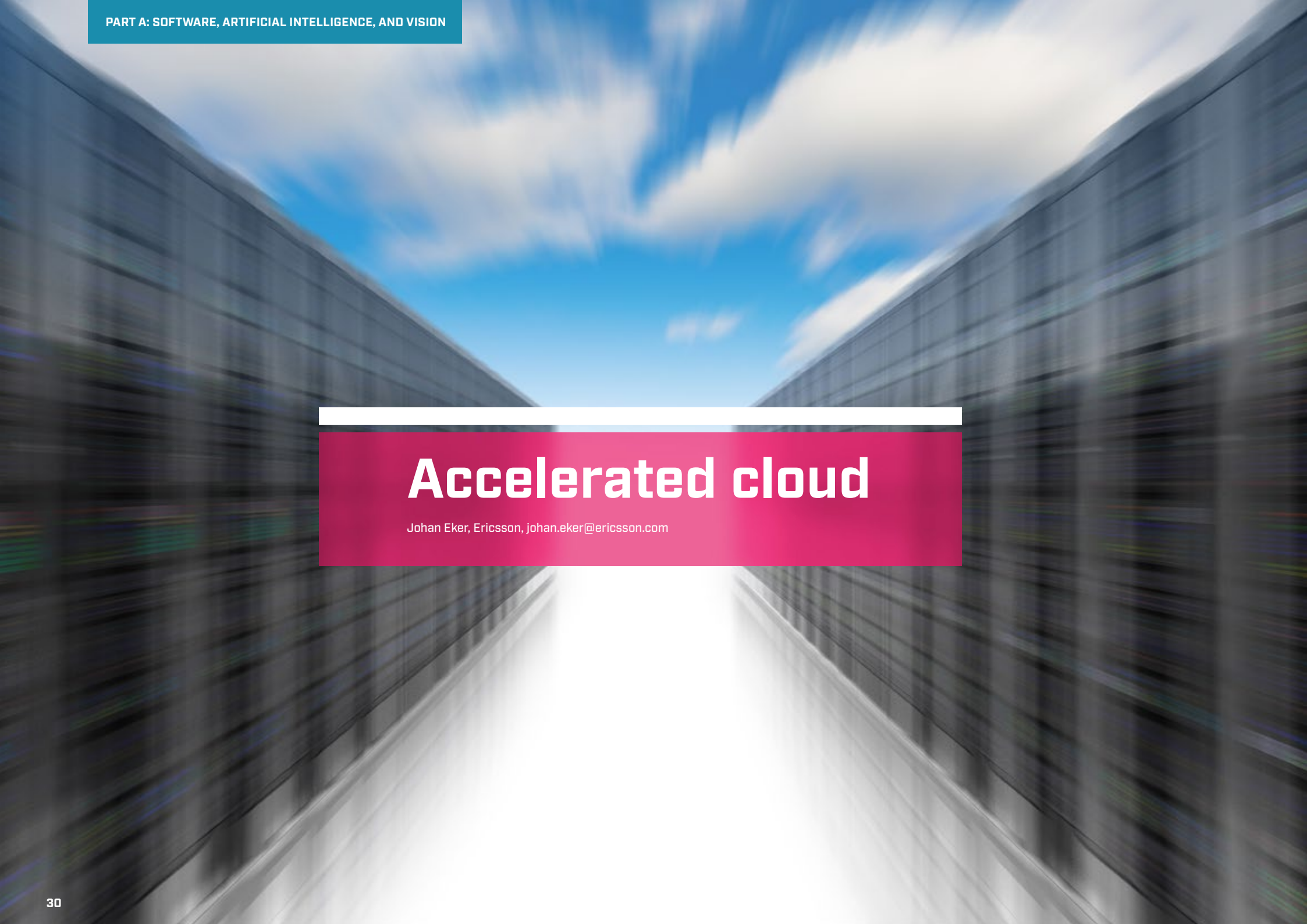
- Oracle has its core team for the development of the Java Virtual Machine (JVM) present in Stockholm. Their expertise in compiler, runtime code modification and monitoring are essential ingredients for diversification and antifragility.
- Ericsson has a strong culture of complex software-development and reliable programming-language design3, which is a major asset to designing and experimenting the programming models that will support developers in the construction of antifragile software.
- Saab has been integrating key open-source technology to build high confidence in large software through advanced software supply chain.
- Spotify is a world-leader in the development and continuous improvement of Internet-scale software. Their expertise in technology to automatically build and deploy provides key insights to build relevant technology for dependability.

The five universities present in WASP have computer-science departments with strong faculty in key areas for the construction of autonomous capabilities in software systems.

NEED FOR SUPPORTING ACTIVITIES

At KTH, there shall be an opening for Professor in operating systems, a key research area to integrate diversification and antifragility at the heart of large scale autonomous software systems.

Footnotes and links:
1. Polyverse: www.polyverse.io
2. Gremlin: blog.gremlininc.com
3. Erlang and the Let-it-crash philosophy: wiki.c2.com/?LetItCrash



Accelerated cloud

Johan Eker, Ericsson, johan.eker@ericsson.com

DEFINITIONS

FPGA: Field-programmable gate array.

ASIC: Application-specific integrated circuit.

GPU: Graphics-processing unit.

API: Application programming interface.

OVERVIEW

Today, cloud computing provides a model for sharing generalized compute resources and data storage connected with millisecond (end-to-end) latency networking. The ability to share computing, storage and networking resources among thousands of potentially competing users, or tenants, is built on top of a platform of virtualization technologies that enable resource isolation between tenants.

Machine learning and image processing benefit vastly from the use of GPUs and show an orders-of-magnitude increase in performance. FPGAs offer more generic accelerator possibilities, at the cost of higher complexity. In the telecom industry there are an ongoing initiatives on virtualization of networking and radio functionality.

Applying the virtualization model to other types of hardware devices besides general-purpose processors and conventional storage is an attractive proposition that is currently just beginning to get started. These devices include FPGAs, dedicated ASICs, GPUs, vector-processing units, and, in the further future, neuromorphic chips and quantum-computing hardware.

INTERNATIONAL MATURITY

Microsoft Research has implemented software framework called Catapult[1], for virtualizing FPGAs. Microsoft is now in the process of deploying this[2] in their Azure Cloud. They also have a cloud GPU service, featuring VMs with NVIDA GPUs, called Azure N Series VM[3]. Amazon has a deployed FPGA service called F1[4], and a cloud GPU service called Elastic GPUs[5]. Google is trialling a cloud service, called Cloud TPU[6], for their TensorFlow AI processor, and a cloud GPU service[7]. IBM is offering cloud access[8] to its quantum computer cloud at no charge to researchers.

Cloud services are characterized by resource pooling, elasticity, self-service and metered usage[9]. Having multiple tenants sharing the same hardware is fundamental to the proposition of cloud. However, providing a cloud service model for accelerators has proven more difficult than for standard CPUs. Accelerators such as GPUs and FPGAs are commonly exposed to VMs (virtual machines) using a so called "pass-through", which means that the VM may access the accelerators as if it was connecting it from a physical machine. This limits resource sharing since the accelerator is locked to one VM even if that VM is dormant.

The services provided today by Amazon and others clearly show the applicability of accelerators in a cloud setting, but are still expensive and difficult to use, which will prevent any large-scale uptake.

Challenges

An important challenge is coming up with system-software abstractions that can be applied to a broad and diverse range of hardware, but nevertheless allow software-application developers to have enough control over the hardware to enable their applications. An example from the operating-systems world is device drivers. Device drivers provide an API towards the operating system that hides much of the complexity of the device and only exposes enough to enable the operating system to control the device. While all the types of hardware described above are very different, they all very likely share aspects of their control and management that would benefit from such abstractions.

These abstractions then need to be incorporated into a comprehensive service platform for application developers that supports applications like machine intelligence and AI.

TYPICAL INHERENT TECHNOLOGIES

Creating virtualization solutions for new hardware accelerators and offering them through a cloud platform is likely to replace a collection of custom, one-off, software solutions that are specific to different kinds of hardware. Such solutions are complex to maintain.

This means that the typical inherent technologies are:

- **Computer architecture:** physical-level access to the accelerators and low-level software designed for efficient communication. Hypervisor for accelerators for providing multi-tenant support.
- **Computer science:** tools and development models for configuring and programming accelerators in a cloud environment.
- **Use-case domains:** machine learning, analytics, control – design of frameworks that utilize cloud accelerators.
- **Service-delivery models** and **domain-specific application frameworks.**

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

Abstractions for particular hardware devices and abstractions across hardware devices are lacking. Development of such abstractions, and their embedding in a body of platform software, is likely to lead to the breakthrough of a well-engineered, common cloud platform for multiple kinds of diverse hardware.

POTENTIAL FOR SWEDISH POSITIONING

Relevant positions would be in the area of electrical engineering, specifically computer hardware/computer science/computer architecture with focus on programmable hardware for cloud. This includes research around GPUs, FPGAs and coarse-grained reconfigurable hardware. The proposed work should mainly focus on the interface rather than the hardware itself. It should have a strong focus on the cloud-delivery model.

POTENTIAL PLAYERS

- Ericsson Research – Cloud System Software Technologies is focusing on flexible software-defined infrastructure including hardware accelerators among other things.
- The groups at Computer Science at Lund University (Jorn Janneck) and Electrical Engineering at Lund University (Fredrik Tufvesson) have long experience with tools for FPGA and accelerator design. In addition, they also have competence in the area of radio accelerators.
- Ericsson is developing and selling a complete cloud stack, including server hardware, management software and applications.
- Michael Felsberg at Linköping University leads an image-processing group that uses accelerated machine learning and could be part of providing use cases and evaluation.
- Axis (Mikael Lindberg) designs and manufactures networked cameras and also provides cloud-based image-processing services, for instance for classification.
- Zenuity are currently building up a machine-learning cluster.

NEED FOR SUPPORTING ACTIVITIES

The area would benefit for access to one or several data centers equipped with relevant hardware. This includes support for racking, configuring and operating the servers and the attached accelerators.

Footnotes and links:

1. www.microsoft.com/en-us/research/project/project-catapult
2. azure.microsoft.com/en-us/resources/videos/build-2017-inside-the-microsoft-fpga-based-configurable-cloud
3. gpu.azure.com
4. aws.amazon.com/ec2/instance-types/f1
5. aws.amazon.com/ec2/Elastic-GPUs
6. cloud.google.com/tpu
7. cloud.google.com/gpu
8. www-03.ibm.com/press/us/en/pressrelease/52403.wss
9. csrc.nist.gov/publications/detail/sp/800-145/final

Cloud

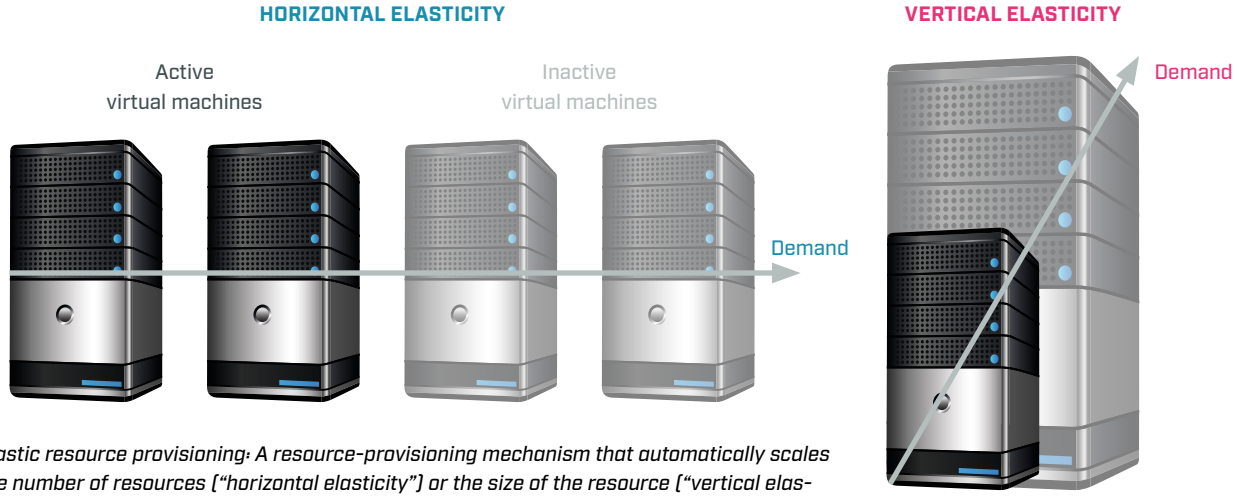
Martin Monperrus, KTH, martin.monperrus@csc.kth.se
Benoît Baudry, KTH, baudry@kth.se

DEFINITIONS

Service-level agreement (SLA): A contract between a service provider and the end user that defines the level of service expected from the service provider.

Disaggregated datacenters: A datacenter that consists of pools of different types of resources, for instance CPU cores, memory, and network, rather than of a set of servers.

DevOps: A software-development and delivery process that emphasizes communication and collaboration between product management, software development, and operations. It supports this by automating and monitoring the process of software integration, testing, deployment, and infrastructure changes by establishing an environment where building, testing, and releasing software can happen rapidly, frequently, and more reliably.



Elastic resource provisioning: A resource-provisioning mechanism that automatically scales the number of resources ("horizontal elasticity") or the size of the resource ("vertical elasticity") with the demand. One example of a resource of this type is a virtual machine (VM). Elasticity is also known as auto-scaling.

OVERVIEW

Cloud technologies are here considered to include a relatively broad set of technologies for elastic on-demand provisioning of ICT resources. The scope includes both public and private cloud realizations; both single datacenters and highly distributed mobile edge clouds extending on telecom networks; both full-scale cloud solutions and critical building blocks such as rackscale systems; as well as a substantial part of the software stack from virtual infrastructures to applications, including the software tools aimed for their design and their operation.

The area is driven by a few independent trends and needs. The rapid increase in use of Internet services leads to enormous and highly varying resource demands. There are high expectations on a rapid growth of Internet-of-Things (IoT) throughout society, leading to yet higher resource demands. As applications go from basic web-content delivery to business-critical enterprise applications, mission-critical applications for system management and control and response-time critical IoT environments including augmented reality, new demanding performance requirements arise.

It is foreseen that today's best-effort clouds based on mega-scale datacenter resources will have to be complemented by highly distributed resources and infrastructure services providing high and guaranteed performance, for instance in terms of throughput and response time. These stronger requirements on high and guaranteed performance are in conflict with increasing needs to substantially improve cost- and energy efficiency. The need to deliver better performance using less resources is a main challenge to be tackled through more sophisticated resource management.

INTERNATIONAL MATURITY

The basic type of cloud offerings, providing pure infrastructure on-demand, made available in pre-defined capacity classes but only with best-effort performance commitments, can today be considered relatively mature. The cloud providers have automated their operation to a great extent, not least to achieve cost reductions in personnel and to enhance system reliability.

This work has also proven rather successful in handling the extreme scale of the largest datacenters. The optimization and autonomous management for performance and for optimizing resource utilization is, however, rather limited. In the research forefront, though, much more efforts have been directed to the performance and utilization issues, but in these cases typically without truly addressing the scale of the large datacenters nor addressing the full complexity of optimizing, for example, performance while simultaneously considering all other relevant optimization objectives and operational requirements.

All-in-all, for taking the final step towards truly self-managing and self-optimizing resource-management systems providing also performance-based service-level agreements (SLAs) for resource provisioning, there are still substantial steps to be taken.

Beside these basic cloud-infrastructure offerings, we here also consider the highly distributed mobile edge clouds and the rackscale systems that are expected to replace traditional clusters as the main building blocks for cloud datacenters. Both these areas are on a much earlier and less mature stage, mainly addressing enabling technologies and early experimental platforms. The efforts on autonomous systems for the operation of these infrastructures are at a very early stage.

Obviously, the industrial research is strong in this area, driven by large companies such as Google, Microsoft, IBM, VMware, Intel, HP and others. In the US, there are many smaller academic research groups with substantial strength on particular subtopics. Among universities with larger strong environments, in particular UC Berkeley should be mentioned.

European industries and academia have difficulties to compete with the strongest US efforts, although the Swedish research in this area is clearly among the strongest in Europe. There are also many, and in some cases large, academic and industrial efforts in Asia, most not yet having very strong impact.

CHALLENGES

The main challenge is inherited from the complexity of a broad range of problems and the fact that they cannot be solved in isolation. For example, an optimized self-management system depends both on the architectures of the infrastructure to be managed and on the programming models defining how applications are interfacing the infrastructure.

Complementing this, there is a broad range of issues regarding how to perform the autonomous management, including when and where to allocate resources for various applications to meet requirements on throughput, response time, cost efficiency, energy efficiency and similar, given that load may vary drastically in both size and locality.

The fact that these very different problems have to be solved in concert is a major part of the overall challenge. That challenge is reinforced by the fact that it is difficult to get access to proper testbeds, realistic in both type and scale, and that workloads of representative type and scale are rare, not to say non-existing, in particular when addressing the use of future infrastructures, including workloads from large-scale internet of things applications or highly demanding rackscale applications.

TYPICAL INHERENT TECHNOLOGIES AND POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

As already indicated, the area is characterized by many hard problems that are closely interrelated and magnified by the system scale. We have already touched upon the need for architectures and programming models for disaggregated datacenters (including rackscale systems) and mobile edge clouds (including network-function virtualization and associated infrastructure for the dynamic provisioning of the network itself).

Sample subproblems in this context are cloud-native programming models and program-development environments, cloud-native distributed operating systems, support for managing the trade-off between data consistency and performance, software design for volatile resources, capabilities to managing the increased resource heterogeneity following from increased use of hardware accelerators, and similar.

The scope for management actions should also go beyond reliability and performance at runtime to more of self-deployment, self-configuration, more sophisticated self-healing and similar.

For many of the management actions, there is a need for better understanding the stochastic properties of the environment, for instance the ability to deduce probabilistic guarantees given stochastic failures.

There is a need to provide more knowledge from the vast amount of monitoring data that potentially can be made available. This calls for additional efforts in machine learning, for instance, using deep neural networks, for example to adopt such methods when solving cloud-management and analysing problems, and to become better in training the neural network in that context. Today's methods are typically considered too slow to handle the extremely large amount of data considered.

There are also situations where improved ability to understand the neural networks would be beneficial (without having the models). When, for instance, performance anomalies are identified, there is a need to understand the root-causes, using formal methods, probabilistic reasoning, network theory or similar, together with improved performance models of the underlying systems.

As the autonomous resource management is not expected to be fully autonomous, but rather to support system administrators with powerful tools allowing for policy steering rather than actuation, it is relevant to further investigate the human-in-the-loop-aspects. This can be extended to also cover the whole software pipeline for systems, including DevOps, automatic testing, regression testing, iterations with feedback from deployment and operation to development and similar.

The edge-cloud scenario will create new problems due to buffer overload, calling for networking research to provide throughput without drastically increasing latency with too many/large buffers. Security requirements will ask for additional efforts and it is anticipated that further enhancement of block-chain and distributed ledgers, for instance with respect to scalability, will be important.

Intel's hardware-assisted security approach also opens up new opportunities to be explored.

POTENTIAL FOR SWEDISH POSITIONING

There are a couple of ongoing and potential initiatives that can be identified as having high potential for Swedish positioning. The Swedish activities often referred to as "cloud control" (associated with the Cloud Control Workshop series[1]) have begun to put Sweden on the map for methodological research for cloud-resource management, combining distributed-systems research with research in not only control theory (as the name suggests) but also discrete optimization, mathematical statistics, artificial intelligence including machine learning and similar. These efforts may be strengthened and further exploited for Swedish positioning.

There is also potential for Swedish positioning in the subarea of cloud control addressing resource management for rackscale systems, taking advantage of both the Ericsson leading initiatives on implementing the Intel Rackscale architecture and to extend the current cloud-control activities with additional efforts on (memory) architectures and programming models.

A third opportunity could be to address the area of management of IoT infrastructures and applications, again extending on cloud-control activities and Ericsson efforts, but here also involving IoT stakeholders in some key application areas, for example for industrial processes, automotive industry or health and well-being. One could argue that the positioning for this opportunity is not as strong as for the two above, but significant advantages may be provided through a leading testbed including both 5G technology and an IoT application platform (for instance Calvin[2]).

POTENTIAL PLAYERS

WASP activities in this area are today mainly driven by Lund University and Umeå University with some involvement from KTH. There is potential to increase the engagement from all of these but also to involve Chalmers and to some extent Linköping University in this work. Additional Swedish academic partners could primarily be Uppsala University (if addressing the rackscale scenario). There are also related activities at Mälardalen University, Karlstad University, Luleå University of Technology and Blekinge University of Technology.

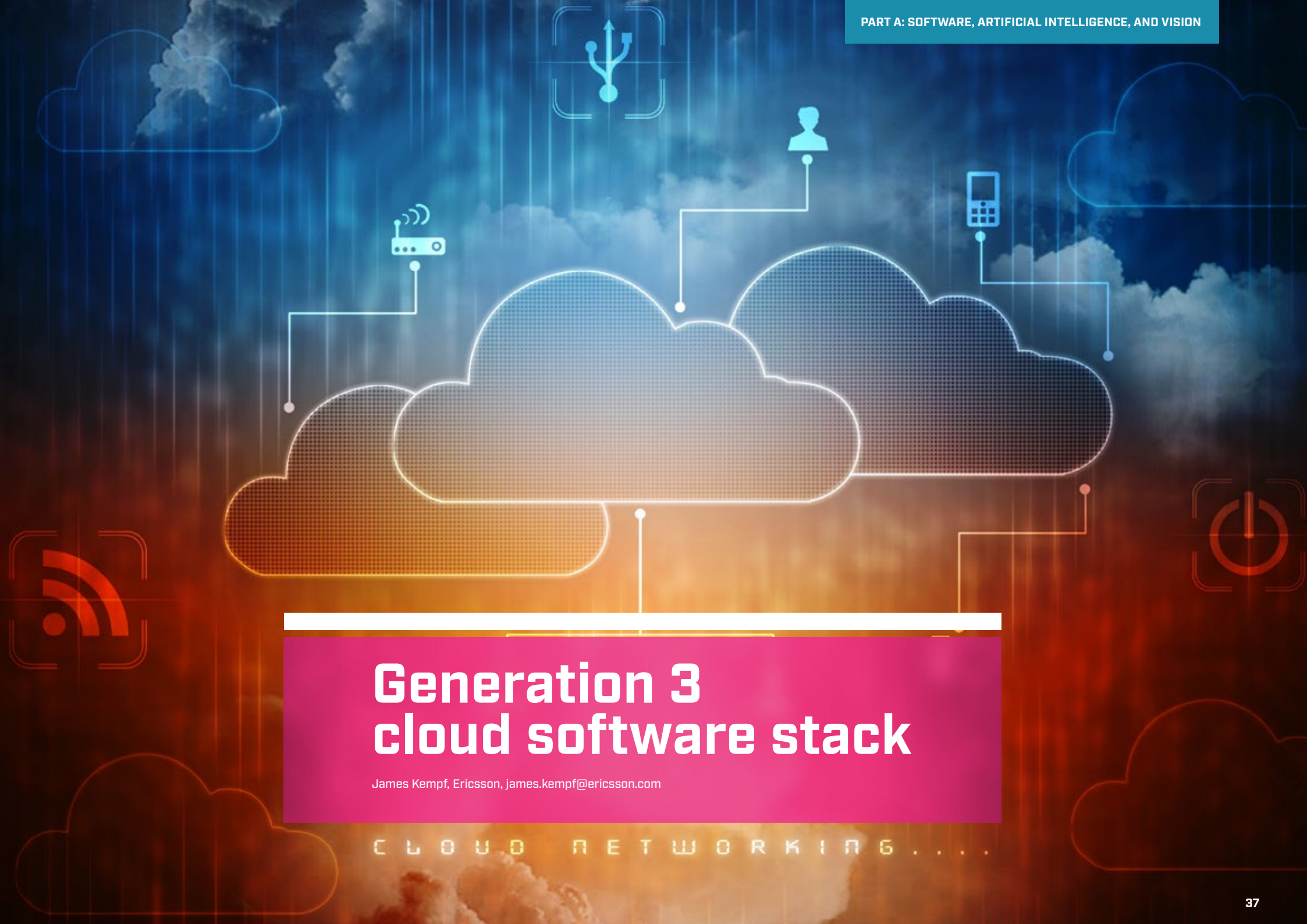
Among industrial players, Ericsson stands out as the most important, but relevant players include also other major IT companies, like Tieto or Swedish subsidiaries of global giants like IBM an Microsoft. Also highly relevant are a large number of companies that may be considered rather as application owners than primarily IT companies, while also doing substantial work in this area in-house, with examples like Volvo, Saab, ABB, Sandvik and SKF, and many smaller companies, including those already involved with industry PhD students within WASP.

NEED FOR SUPPORTING ACTIVITIES

Academic research is this area has a major disadvantage compared to industry giants as Google, Microsoft, IBM and similar, by not having access to the real and large-scale workloads and testbeds, as well as the hands-on experience from running these. If Sweden could provide realistic large-scale testbeds and realistic and large-scale workloads (real or synthetic), this would provide Swedish research with an advantage over basically all academic research in this field.

Something else that could substantially have large impact are large-scale joint efforts on relatively basic research in this area. EU FP7 and H2020 have funded several large-scale projects in the field, but typically being relatively broad and driven by specific application needs. A large concerted effort on more basic research on autonomous management could potentially be very successful.

Footnotes and links:
1. www.cloudresearch.org/workshops
2. Calvin is an open-source Internet-of-Things application-development framework developed by Ericsson Research, www.ericsson.com/research-blog/open-source-calvin



Generation 3 cloud software stack

James Kempf, Ericsson, james.kempf@ericsson.com

DEFINITIONS

RDMA: Remote Direct Memory Access.

DRAM: Dynamic Random Access Memory.

OVERVIEW

Some important characterteristics of any cloud platform are elastic resources, multi tenancy and resource pooling. This is, with few exceptions, achieved through the use of virtual machines with some additional support software stack.

Today's platform software is fragmented between laptops and data centers, complicating the developer's interaction with the compute environment. Most developers do their development on laptops, then write infrastructure code to deploy their applications onto various cloud platform. Each of the major public cloud platforms has its own proprietary way of deploying applications, requiring developers to maintain separate code bases for infrastructure deployment programming in different public clouds.

An ideal software development platform would allow developers to develop in a cloud-native fashion, in exactly the same manner as the cloud-native deployment tools that have appeared over the last few years. This platform, generically called Gen3 Cloud, has the following characteristics:

- The main execution abstraction is not virtual machines or containers, but instead provides a more fine-granular model that is detached from the underlying hardware infrastructure. The application models are native to the cloud, meaning that they are designed primarily for the cloud and not retrofitted. Emerging patterns that should be supported are microservices or function as a service (FaaS), also known as serverless programming.
- Deployment should be deterministic and greatly simplified compared to the orchestration systems of today. Applications are developed either directly in the cloud or on the developer's laptop, with deployment requiring only a press of the (metaphorical) "cloud button".
- The platform offers an ecosystem so developers can offer their products to others and get paid for their use.
- The platform enforces selective and fine-grained authorization combined with identity management so developers can choose who gets access to their products.
- End-to-end networking latency drops from 1 millisecond to 10 microseconds, so networking, effectively, disappears within the data center so that developers need not pay any more attention to it than they currently pay to the bus between memory and the CPU on a server.
- Storage latency drops to within a factor of DRAM (main memory) latency via non-volatile memory so that shared key-value stores become transparent.

These characteristics require the development of new platform software, language runtime systems, and API abstractions.

INTERNATIONAL MATURITY

Bits and pieces of the Gen3 Cloud vision are starting to appear. Google[1] is working on networking to enable transparent, ultra-low latency network. The Cloud Native Computing Foundation[2] is working on platform software for transparent microservice deployment. Open source efforts such as Istio[3] are beginning to address the prerequisites for a microservice ecosystem.

This work is quite fragmented and doesn't address the platform or system problem. Work in the late 1980s and early 1990s on distributed operating systems, such as Ameoba[4], is perhaps more relevant. This work enabled an operating system to span across multiple servers within a local area network. The drawback was that, at the time, processor performance was scaling much more quickly than networking performance, so operations such as distributed shared virtual memory were too slow. Recently, some architectural thinking about modifying language runtimes[5] for cloud environments has begun to appear.

CHALLENGES

There are a whole list of challenges around how current cloud platforms structure developer access to compute networking and storage that need addressing to make this vision possible.

A major challenge is coming up with a collection of communication abstractions that are grouped by average and tail latency, so that developers can choose how to provide access to their applications depending on the communication distance, and provide program development tools for these abstractions. This has relevance for distributed/edge cloud as well.

Another challenge is how to incorporate non-volatile memory in a way that makes programmatic access easy. Implementing a distributed language runtime is an important challenge. Yet another challenge is reengineering the server operating system so that it more resembles a distributed operating system, with resource allocation and placement spanning across the data center.

Finally, a programming model that natively supports the features in the cloud, such as elasticity and distribution, which defines and encapsulates important programming patterns. Futhermore, deployement should be greatly simplified.

Integrating the above features into a usable platform for developers, spanning laptops to data centers, will require a substantial amount of system-design research.

TYPICAL INHERENT TECHNOLOGIES

Most data-center-management systems today are centralized and offer no support for programmatic abstractions that take cloud characteristica, for instance distribution,

elasticity, and latency into account. If and when Gen 3 Cloud happens, it will radically alter how developers develop and deploy software, just as Gen 2 Cloud did in the recent past. Mostly the deployment step will disappear.

Relevant research areas are:

- **Operating system research** – design of a new cloud platform that supports a distributed, scalable, elastic, failure-resilient, and latency-aware execution model and that supports the emerging design patterns and next-generation data-center hardware.
- **Software engineering** – new tools, languages, and developer processes to improve developer productivity and application quality. This includes programming models and developer tools, such as profilers and debuggers, as well as software-engineering processes.
- **Computer Networking** – supporting software for multi-tenant networking on network substrates that exhibit radically reduced network latency will be necessary. Tenant isolation on low-latency/high-speed network substrates is currently an unsolved problem. More fine-grained tenant isolation for applications that communicate within the data center is one possible approach. Network-management software for low-latency/high-speed networking is also lacking. Application of machine intelligence, AI, and analytics to network management within the data center, as well as between data centers for a distributed cloud, are required to reduce the human effort and possibility of errors, which can lead to costly downtime.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

For low-latency networking there are some technologies available, such as RDMA, but they are still not widely deployed and their multitenant isolation and network management support is poor. Abstractions for programmers that are structured along latency lines are needed. System software for NVM access is needed. Multitenant isolation techniques that are lightweight (meaning that they don't require cryptography or network tunnels) are required. Finally, integrating the whole bundle into a usable platform is absolutely required, otherwise it is not possible to determine whether the individual technologies are usable by anyone.

POTENTIAL FOR SWEDISH POSITIONING

Cloud technology will be an inherent part of a future where many functions, that today are carried out by humans, are being automated or made autonomous. Simplfying development and operations will allow for non-IT companies to enter the cloud. Improving performance will lead to lower cost and new application domains. Sweden needs to build and host its own cloud infrastructure.

Relevant areas for research positions:

- Computer science – language design, compilers, debuggers, profiling.
- Distributed computing – efficient, low-latency communication in distributed systems.
- Control system – use cases, driver applications.

POTENTIAL PLAYERS

- Ericsson Research – research area Cloud System Software Technology is involved in all parts of the cloud software and also operates a research data center.
- Volvo and Autoliv provide new and challenging use cases.
- The University groups at KTH (Dejan Kistic), Umeå University (Erik Elmroth) and Lund University (Karl-Erik Årzén) are instrumental for cloud and use-case competence.

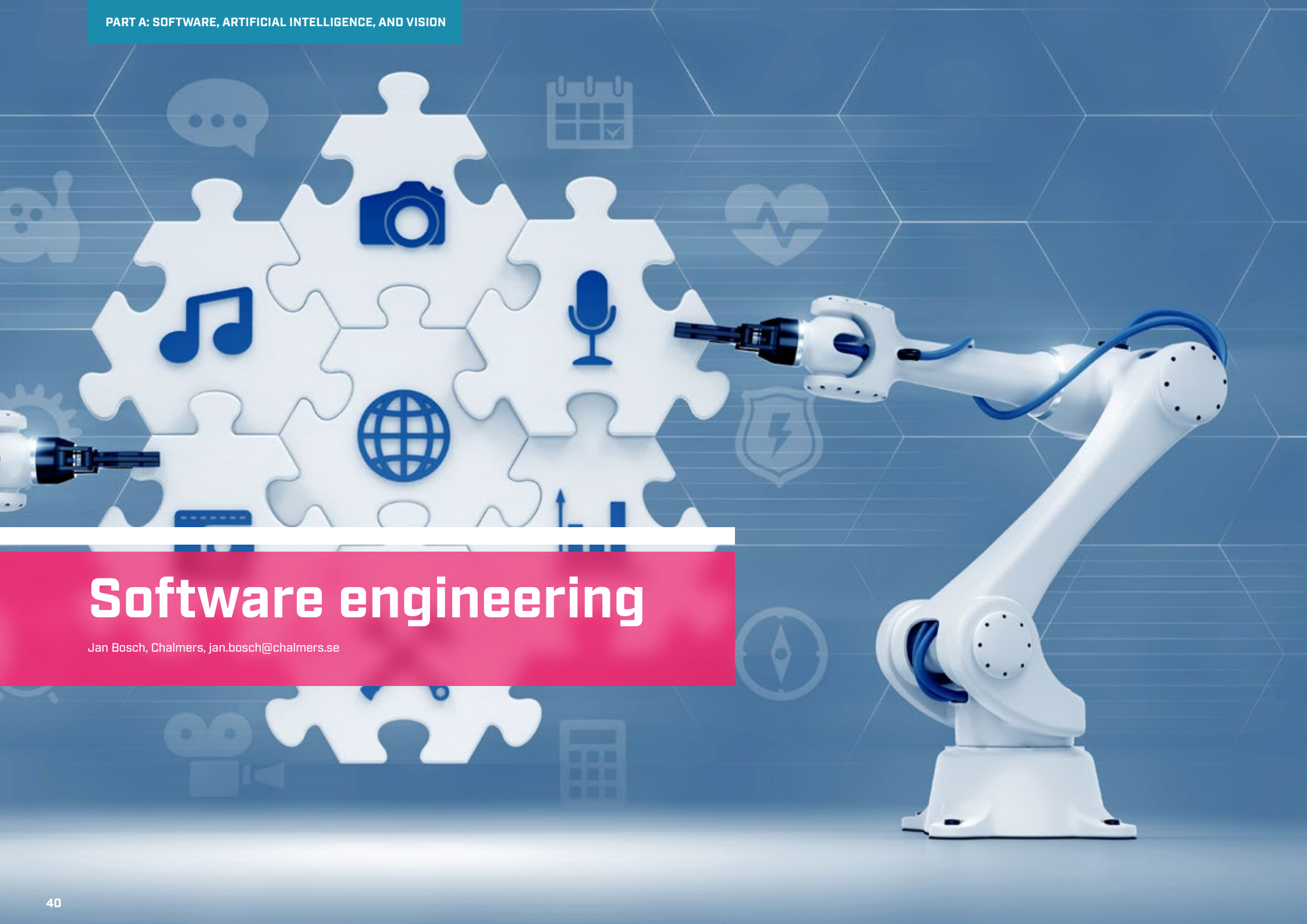
NEED FOR SUPPORTING ACTIVITIES

Cloud research is by and large an implementation-oriented and hands-on operational activity. A major effort to increase substantially beyond the state-of-the-art will require some considerable implementation of concepts and put them in operational environment for further research on potential enhancement based on active workloads.

Since the latter is difficult to perform purely in an academic setup, it will benefit greatly in a WASP context. The data center run by Ericsson Research – research area Cloud System Software Technology in Lund provides a suitable cloud infrastructure for research effort and operational experimentation for the WASP program.

Footnotes and links:

1. events.linuxfoundation.org/sites/events/files/slides/ONS Keynote Vahdat 2017.pdf
2. www.cncf.io
3. www.istio.io
4. ieeexplore.ieee.org/document/53354
5. pdfs.semanticscholar.org/4b4c/6dad77e546b4ff3ebdb27f2ea419445631b9.pdf



Software engineering

Jan Bosch, Chalmers, jan.bosch@chalmers.se

DEFINITIONS

CMMi: Capability Maturity Model Integration, a process-level-improvement training and appraisal program.

SaaS: Software as a service, a software-licensing and delivery model in which software is licensed on a subscription basis and is centrally hosted.

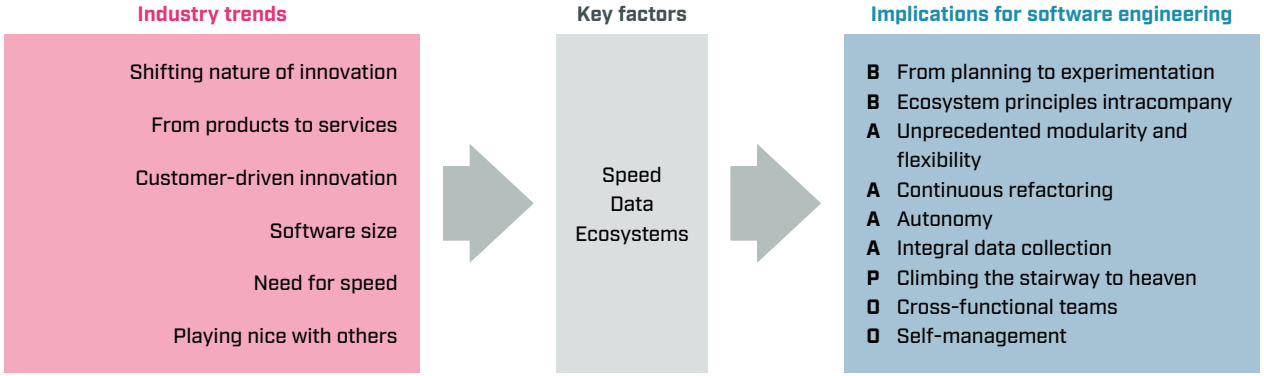


*Figure 1: The trends and factors influencing software engineering's future, and the implications for **B** business, **A** architecture, **P** process, and **O** organization.*

OVERVIEW

An evaluation of recent industrial and societal trends revealed three key factors driving software engineering's future: speed, data and ecosystems. These factors' implications have led to guidelines for companies to evolve their software engineering practices.

Industry investment in software R&D is increasing[1], and software, rather than mechanics and hardware, now defines a product's value[2,3]. So, industry is under severe pressure to improve software-intensive systems' capabilities to deliver on today's software needs.

INTERNATIONAL MATURITY

The future is generally difficult to predict. However, several societal and technological trends indicate that certain developments and transitions will occur, even if it's not clear when. Here, I summarize six trends influencing the evolution of software engineering practices. The trends are organized from high level and societal to more specific and technological.

The Shifting Nature of Product Innovation

In the past, especially in the embedded-systems industry, a system's or product's mechanical parts were most often targeted for innovation. By introducing new materials, presenting alternative designs that reduced weight or increased structural integrity, or adhering to fashionable designs, companies could differentiate their products. Even if the product contained electronics and software, these technologies were considered secondary and not necessarily central to the product. The software had to work but wasn't viewed as differentiating for the product.

In addition, software development was subjugated to mechanical development, even if the software could be developed significantly faster than the mechanical system.

The trend. Now, software is becoming the central differentiator for many products, whereas mechanics and hardware (electronics) are rapidly becoming commodities. In addition, the system architecture often seeks to separate the mechanics and hardware from the software to allow for two largely independent release processes. So, software can be updated frequently, both before the product leaves the factory and after it's been deployed to customers. As part of this trend, customers increasingly expect their product's software to evolve.

The evidence. At the Software Center (see info box), several companies have undergone this transformation. For instance, AB Volvo estimates that software drives 70 percent of all innovation in its trucks. Volvo Cars estimates that electronics and software drive 80 to 90 percent of its innovation. Over the last decade, telecom company Ericsson's focus has also shifted, with more than 80 percent of its R&D budget dedicated to software. A recent Harvard Business Review article confirmed this trend[4], as did Valeriy Vyatkin's state-of-theart review showing that the ratio of software in machinery has doubled from 20 to 40 percent over the last decade3.

From Products to Services

Businesses and consumers are increasingly aware of capital expenditures' limiting effects. Owning large, expensive items, often funded by borrowed capital, is expensive and limits a company's ability to rapidly change course when customers demand changes. So, many companies are moving from owning buildings, equipment, and other

SOFTWARE CENTER

Software Center is a partner-funded collaboration founded in 2011 focused on accelerating the adoption of best practices concerning large-scale software engineering.

The industrial partners include Ericsson, Volvo Car Corporation, Volvo Group, Saab Defense, Tetra Pak, Axis Communications, Grundfos, Jeppesen (part of Boeing), Siemens and Robert Bosch. The academic partners include Chalmers University of Technology, University of Gothenburg, Malmö University, Linköping University and Mälardalen University.

The center funds research projects in close collaboration with industry on topics including continuous delivery, software architecture, metrics, customer data and software ecosystems.

The mission of Software Center is to help companies to deliver fast and continuous value to their customers and its success factors are equally balanced between academic excellence and industrial impact.

Since its founding, Software Center has grown from six to 16 partners, from three to more than 20 projects and from four to close to 30 researchers.

capital-intensive items to service arrangements in which they pay a fee to access the facility or item.

Consumers, especially in this age of Generation Y, are also shifting their values from owning to having access to expensive items. Developments such as the access economy exploit the fact that many people own expensive items but use them for only small amounts of time each day or week. For instance, the typical car is used less than an hour per day.

The trend. Many industries, including automotive and telecommunications, are fundamentally changing their business models and thus companies' key incentives. This move from products to services has two implications. Firstly, the focus changes from selling as much of a product as possible to providing as many services as possible at the accepted quality level. For example, for a car company beginning to provide mobility services, the question becomes how to provide them with as few cars as possible because the product is now a cost item. Secondly, companies have an incentive to maximize their products' economic lives. For example, companies often deploy new software in products already in the field because that approach is the most cost-effective.

The evidence. Besides academic sources[1,5,6], two industrial examples illustrate the trend. Firstly, Ericsson's global services unit is growing faster than its product units, in terms of revenue and staff[7]. Operators look to access their network as a service, focusing on customer acquisition and market share growth. Secondly, automotive companies expect that by 2020, between a third and half of their cars will be used through service agreements rather than ownership.

From Technology- to Customer-Driven Innovation

Technology forms the foundation for innovation. New technologies enable new use cases and let consumers accomplish their goals in novel ways. So, companies invest heavily in technology innovation with the expectation of being rewarded with product differentiation that drives sales and sustains margins.

However, for several industries, despite using patents and other IP protection mechanisms, new technologies tend to become available to all players at roughly the same time, as I mentioned before. So, these companies derive little benefit in terms of differentiation. As technology-driven innovation's benefits decrease, companies increasingly prioritize customer-driven innovation[8].

The trend. Customer-driven innovation involves identifying and meeting new customer needs as well as better meeting known customer needs. This requires deep customer engagement in both qualitative and quantitative ways. Instrumenting software systems, both online and offline, to collect customer behavior data is critical for customer-driven innovation because successful innovations are often developed before customers even express the needs the innovation addresses[8].

The evidence. Petra Bosch-Sijtsema and I studied companies that had adopted new techniques to collect customer insight as part of product development8. This trend is also evident in how new industry entrants have disrupted or are disrupting incumbents. From stalwarts such as Amazon for retail and Tesla for automobiles, to Uber for taxi transportation and Airbnb for hospitality, none of these companies disrupted or won in their markets by using better technology than the incumbents. Because the incumbents better understood their customer bases and had vastly more resources, they might have used technology to address customer needs as well as or better than these new entrants. However, the new entrants better understood customers' unexpressed needs and developed innovative approaches to meet them.

Software Size

For product innovation to move from mechanics and hardware to software, new features and functionality must be realized through software rather than other technologies. This has obvious implications for software size, relative and absolute R&D investment in software, and other product development aspects.

The trend. Depending on the industry, software's size in software-intensive systems is increasing on an order of magnitude every five to ten years. Industry often underestimates this growth's implications. The main challenge is that a software system ten times larger than the previ-

ous generation's requires new architectural approaches, different ways to organize development and significant modularization of testing, release, and postdeployment upgrades. This growth also incurs the complications of running a larger R&D organization. To address these challenges, companies employ approaches such as modular architectures, IT services, and open-source components.

The evidence. Several studies have documented software growth in software-intensive systems. One of the most illustrative studies is by Christof Ebert and Capers Jones, who analyzed this trend for embedded systems[2]. Vyatkin came to many of the same conclusions[3].

The Need for Speed

User adoption of new technologies, products, and solutions is continuously accelerating. Once measured in years, user adoption has decreased to months and now weeks and days over the last decade. For example, whereas Facebook took ten months to reach a million users, the Draw Something mobile app reached a million users in just days. With enterprise's "consumerization," corporations are also demonstrating this need for speed, driving toward faster adoption of new applications, technologies and systems.

The trend. Companies today must respond to new customer needs and requests at unprecedented speeds, which requires a level of enterprise-wide agility that's often exceedingly difficult in traditional, hierarchical organizations. The need for speed requires companies to pursue different ways to organize, build and architect software and software development.

Particularly in heavily regulated industries, incumbents often control their products in ways that don't support agility and speed but that slow everything down. New competitors enter these industries from the side and work in relatively unregulated areas, which lets them innovate much more quickly than their incumbent counterparts. Even when compliance is required, new entrants – lacking the existing players' legacy – tend to find more resource-efficient and faster ways to comply.

The evidence. To describe this trend, Larry Downes and Paul Nunes used the compelling phrase "big bang disruption[9]." They presented several cases from various industries in which fast-moving new entrants outcompeted incumbents on price, performance and user experience.

Playing Nice with Others

No company operates in a vacuum, but many large organizations' internal operations receive orders of magnitude more attention than events outside the company. However, this is changing rapidly in many industries with software-intensive systems. Companies are increasingly realizing the benefits of playing nice with others and availing themselves of the opportunities presented by using their partner ecosystem more proactively and intentionally.

The trend. The competitive battleground for companies is shifting from focusing on internal scale, efficiency and quality and serving customers in a one-to-one relationship, to creating and contributing to an ecosystem of players that can include suppliers, complementors, customers and, potentially, competitors. We see the ecosystem trend not only in the mobile industry's app stores but also in business-to-business markets such as those surrounding SAP and Microsoft Office. Establishing and evolving ecosystems of different partner types is the key differentiator in several industries and might ultimately decide which companies win a market and which get relegated to less dominant positions.

The evidence. Bosch-Sijtsema and I discussed several cases in which companies improved their competitiveness by effectively using their ecosystems[10]. However, one of the most illustrative cases is Apple. By creating its App Store, the company established itself as the dominant player in the mobile industry[11]. While competitors such as Nokia were focusing on device quality, Apple was creating a partner ecosystem to build new iPhone applications – and this became a key differentiator for the company.

CHALLENGES

The following three factors are at the heart of the trends I just described.

Speed

Analyses show that the ability to respond quickly to events such as customer requests, changing market priorities or new competitors is critical to continued success. Companies must respond at a constantly accelerating rate, so speed will affect the entire organization, from its business models to its organizational structures.

Data

With storage costs falling to zero and virtually every product's connectivity exploding, collecting data from products in the field, customers and other sources is a reality that's still unfolding. However, the challenge isn't the big data but the organization's ability to make smart, timely decisions based on the data. Although many companies still rely on their managers' opinions, future organizations will increasingly use data to inform decision-making. So, data collection, data analysis, and decision-making based on that data will strongly affect companies' functions, architecture and ways of working.

Ecosystems

Future organizations will have increasingly interdependent ecosystems. Because the ecosystem is central, business success requires intentional, not ad-hoc, management of ecosystem partners. This is true for both large keystone players and the typically smaller complementors. As a result of increased speed and data, companies will have to frequently and aggressively change their role and position in their ecosystems. To effectively manage changing relationships – while forward integrating in the value chain by offering solutions or services and moving backward by providing components or entering adjacent markets – organizations will have to proactively manage the ecosystem.

This will strongly affect software architecture, interfaces and ways of working with partner R&D teams.

TYPICAL INHERENT TECHNOLOGIES

I discuss here the trends' implications for software engineering's future, using the BAPO (business, architecture, process, organization) framework[12].

Business

This area involves two implications: the transition from planning to experimentation and the adoption of ecosystem principles.

From planning to experimentation

Companies must transition from working with planned releases with detailed requirement specifications to continuously experimenting with customers – for example, by optimizing previously implemented features, iteratively developing new features or building entirely new products.

This transition is critical for two reasons. Firstly, research has shown that more than half the features in a typical software-intensive system are never or hardly ever used[13]. Building slices of features and then measuring changes in customer or system behavior is a structured way to minimize investment in features that don't add value. Second, as I discussed earlier, customer needs and desires change rapidly. Companies that don't constantly test new ideas with customers risk being disrupted by companies that more readily identify shifts in customer preference.

Adoption of an experimental approach requires a transition from requirements engineering to hypothesis and data-driven development approaches. Hence, outcome-oriented engineering methods should be prioritized.

Adopting ecosystem principles

Customer experimentation requires organizing in fundamentally different ways. Traditional functions and hierarchies are no longer sufficiently fast and efficient. Teams will require more autonomy to make decisions locally on the basis of qualitative and quantitative data from systems in the field.

Moreover, the sheer size of the systems being built these days makes it increasingly difficult to handle their complexities. Instead, we must view them as ecosystems with several parts and autonomous organizational units responsible for the parts.

This autonomy's principles are similar to those of software ecosystems in which the parties make decisions independently – within the underlying constraints of the ecosystem's architecture and platform – while contributing to the ecosystem's overall goal. Traditional organizations focus on power hierarchies to centralize decision making. Going forward, teams will be increasingly autonomous, and organizational leaders will need to emphasize purpose and culture, which will provide the guardrails for the teams to operate within. In effect, software ecosystem principles will be adopted inside organizations.

Intra-organizational empowerment, software ecosystems and systems of systems share the departure from centralized locus of control to decentralized forms of alignment and governance (a la Blockchain). This means that traditional process-centric forms of coordinating work, such as CMMi, need to be replaced with architecture-centric and principles ways of alignment. This has significant implications on the architectural approach (see below), but also on the business models as traditional client-supplier models are increasingly replaced with other business models such as revenue share, platform fee and FLOSS (Free/Libre Open Source Software) models.

Architecture

This area involves four implications: unprecedented architecture modularity and flexibility, continuous refactoring, autonomy, and integral data collection.

Unprecedented architecture modularity and flexibility

Although modularity and flexibility have been important software architecture elements since their conception, they're often compromised to accomplish operational (runtime) quality attributes such as performance. However, with the increasing importance of speed, experimentation, and team autonomy, modularity and flexibility are being prioritized over other quality attributes.

The microservices architecture that Amazon, Netflix and others employ is an example of a highly modular architecture. In this architectural style, large complex systems are modeled as collections of small, independent communicating processes. Although controlling and predicting architecture properties might seem difficult, doing so provides high flexibility and modularity and easy monitoring of system behavior. The behavior is predictable for similar system loads, which allows comparisons to earlier system executions.

From a technological perspective, the challenge is facilitating architecture-centric coordination across systems, ecosystems and systems of systems. For instance, many companies have IoT deployments from different vendors. The interoperability between the sensors and actuators from these vendors is non-existent today due to lack of suitable business models, but also because of a lack of architectural approaches that allow for interoperability while maintaining security and safety requirements.

Continuous refactoring

Especially in the embedded-systems industry, there's a tendency to treat software like mechanical design – that is, design once and use and extend for a long time afterward.

For software, this leads to the accumulation of architecture technical debt.

Continuous refactoring of software-intensive systems' architecture will maintain the architecture's suitability for its intended purpose and minimize the cost of adding features and use cases. However, architects will have to be able to constantly identify and prioritize refactoring items to use the allocated resources optimally.

Our research shows that systems exhibit technical debt from their inception. This means that for long-lived systems, we need to continuously invest resources in refactoring the architecture of the system. Although the topic of architecture technical debt management has received attention in the research community, more is required.

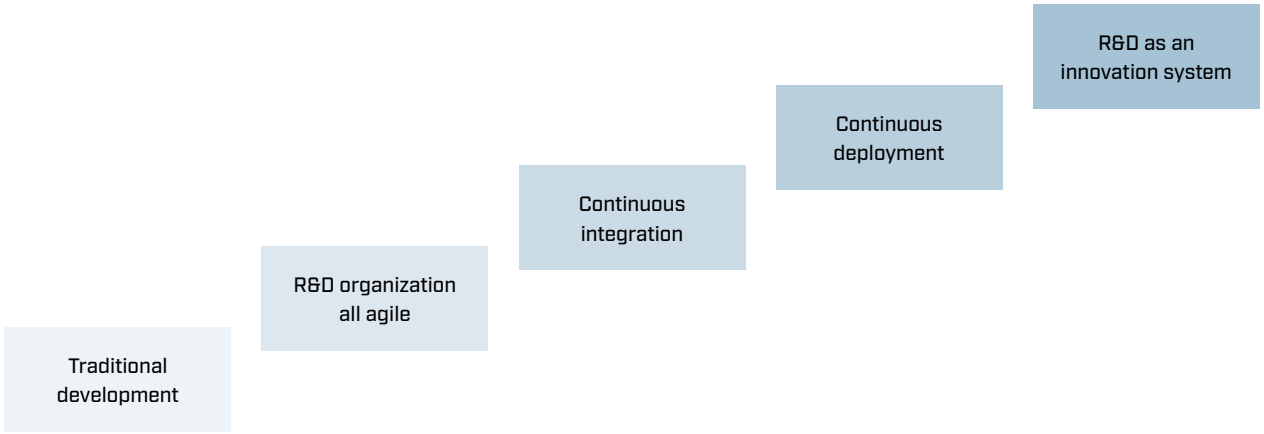


Figure 2: The "stairway to heaven" model describes the evolution of companies' software development processes.

Autonomy

Driven by the transition to services, software's growing size and human labor's high cost, software-intensive systems will be increasingly autonomous. One of the most illustrative examples is the rapid emergence of semi-autonomous cars. Most industries have significant opportunities to transition from semi-autonomous systems that help users accomplish their business goals to systems that autonomously accomplish those goals.

Architecturally, autonomy requires reflective functionality: the system collects data about its performance and adjusts that performance according to its goals. Because data from different parts of the system must be combined to derive relevant information for decision making, architectures will include data-fusion functionality.

Integral data collection

Increasingly, autonomous software-intensive systems need continuous data collection about their operation so that they can control and change their behavior when needed. In addition, feedback from deployed systems and their users is becoming increasingly important for experimentation. So, collecting operational, usage and other data is rapidly becoming integral to architecture.

Future architectures must assume that data collection at all system levels is required and should be integrated by default. Not collecting data for certain system parts will require an explicit decision. As I mentioned earlier, fusing, aggregating and abstracting data will be key requirements for architectures. Such capabilities will be required both for the system's reflective functionality and for informing the system's R&D organization.

The two implications of autonomy and integral data collection provide the perfect preconditions for machine learning and deep-learning solutions. Although these technologies have matured to the point of seeing deployment in a variety of industry contexts, the software engineering implications of building deep-learning systems as well as the engineering implications of systems combining deep-learning components with traditionally engineered software components is poorly understood.

Process

My colleagues and I developed a "stairway to heaven" model describing how companies evolve their development processes from a traditional waterfall style to agile development (see Figure 2)[14]. In that model, companies adopt continuous integration as a core enabling technology. Once new functionality is constantly developed and available at production quality, owing to the continuous integration environment, customers will want to access new functionality before the regular release process. At this point, the company moves toward continuous deployment. Once continuous deployment fully rolls out, the company can run more experiments with customers and the systems installed in the field.

Each step in the stairway has significant implications for work processes, organizational units, tooling and general work methods. In most companies, the steps become increasingly challenging as the required changes involve larger and larger parts of the organization. For instance, besides R&D personnel, the verification-and-validation, release, customer-documentation, customer-support and sales-and-marketing teams must be involved to manage

this fundamental shift in software deployment. Aligning all these groups in a well-functioning process is much more difficult than adopting agile development because it requires changes in the R&D organization.

Organization

The two implications here are cross-functional teams and selfmanagement.

Cross-functional teams. Traditional organizations rely on functionally organized hierarchies that group people with similar skill sets – for example, product management, development, verification or release. Although this allows for pooling of skills and flexible resource allocation to activities, it often leads to slow decision making and execution because of the many handovers between functions and decisions that must go up and down the hierarchy.

Going forward, cross-functional teams will be empowered to make decisions and work with limited coordination between teams. Agile R&D teams are an example. As organizations climb the stairway to heaven, these teams will replace the hierarchical functions. For instance, besides engineers, teams will include members with skills in verification and validation, release, product management and, potentially, sales, marketing and general business.

Self-management. A disadvantage of hierarchical management is the time required to make decisions. In fast-moving, highly complex environments, relying on a hierarchical model is a recipe for disaster. The alternative is to decentralize management to the point that individuals and teams manage themselves. Agile teams today often have significant autonomy. As teams become increasingly cross-functional, selfmanagement will be required to maintain competitiveness. Management will be more concerned with growing and steering the organization's culture, resulting in individuals and teams making good decisions despite the lack of the traditional hierarchies.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

Based on the aforementioned trends, it is necessary that the following technology areas are developed further:

- Software-engineering approaches for building machine-/deep-learning approaches. Current deep-learning deployments tend to be prototypical and not used in production environments, except for a few companies such as Google, Facebook and Microsoft. Maturing the engineering approaches to support this is critical for the Swedish industry.
- Engineering approaches to effectively integrate machine-/deep-learning components in systems that are engineered using traditional software are required as this is currently poorly understood.
- Transitioning to a data-driven and hypothesis/experiment-based approach to engineering, product development, running a business, etcetera is critical. Modern companies such as Microsoft and Booking.com run thousands of A/B experiments in parallel, continuously and use data for every decision taken across the entire company (and not just in R&D). The Swedish industry needs to adopt these principles.
- Effective engineering approaches for cross-organizational, ecosystem and system-of-systems contexts where control is by necessity decentralized is critical in a world of IoT deployments, autonomous systems entering and leaving contexts and companies and individuals increasingly dependent on dynamically composed software systems. These approaches need to ensure security and safety requirements while allowing for unprecendented flexibility.
- Continuous deployment of functionality is a basic enabler for the aforementioned technologies and although the technology is well understood for SaaS deployments, there are many challenges left when it comes to continuous deployment in the context of embedded systems for which dozens, hundreds or millions of instances exist out in the field. This is exacerbated by the need to support this for safety-critical, secure and certified systems.

POTENTIAL FOR SWEDISH POSITIONING

Despite its small size, Sweden is world-class when it comes to building complex, advanced embedded software-intensive systems. As the trend is shifting from standalone systems to dynamically composing systems, ecosystems and systems of systems, the Swedish industry has to build a new set of core capabilities, technologies and solutions. Doing so successfully and earlier than competitors from other countries will position the Swedish industry in a superior competitive position.

POTENTIAL PLAYERS

The challenge for the Swedish industry is less concerned with creation of fundamentally new technologies, but rather with the acceleration of adoption of these new technologies in the existing industry as well as in the startup community. Large companies are notoriously bad at rapid adoption of new technologies, largely driven by what some refer to as the permafrost, which means the middle management layer between front-line engineers and C-suite executives. Initiatives such as Software Center focus explicitly on accelerating the adoption of new technologies, rather than on creating fundamentally new technologies.

Outside Sweden, the focus for collaboration should be on European collaborations as the major risk for disruption of the Swedish and European software-intensive systems industry can be found in the innovative power of Silicon Valley and the cost advantages of Asian competitors. Hence, collaborations with, for example, Fraunhofer IESE, Paluno at the University of Duisburg/Essen, the high-tech campus in Eindhoven and LERO in Ireland should be prioritized.

NEED FOR SUPPORTING ACTIVITIES

Finally, although WASP is rightfully concerned with new and breakthrough technologies, my main concern is that we spend too little effort and time on accelerating the adoption of these technologies in the existing and

new industries in Sweden. Having worked for Nokia and seeing all the technologies that were later used to disrupt the company being present in our research labs, I have experienced up close and personal the challenge of driving adoption of new technologies across the organization.

My recommendation would be to supplement the existing technology research with activities aimed specifically at accelerating the adoption of these technologies at the WASP companies and the Swedish industry at large.

Footnotes and links:

1. M.A. Cusumano, "The Changing Software Business: Moving from Product to Services," Computer, vol. 41, no. 1, 2008, pp. 20–27.
2. C. Ebert and C. Jones, "Embedded Software: Facts, Figures, and Future," Computer, vol. 42, no. 4, 2009, pp. 42–52.
3. V. Vyatkin, "Software Engineering in Industrial Automation: State-of-the- Art Review," IEEE Trans. Industrial Informatics, vol. 9, no. 3, 2013, pp. 1234–1249.
4. W.C. Shih, "Does Hardware Even Matter Anymore?," Harvard Business Rev., 9 June 2105; hbr.org /2015/06/does-hardware-even -matter-anymore
5. H. Gebauer and T. Friedli, "Behavioral Implications of the Transition Process from Products to Services," J. Business and Industrial Marketing, vol. 20, no. 2, 2005, pp. 70–78.
6. R. Oliva and R. Kallenberg, "Managing the Transition from Products to Services," Int'l J. Service Industry Management, vol. 14, no. 2, 2003, pp. 160–172.
7. S.W. Elfving and N. Urquhart, "Servitization Challenges within Telecommunications: From Serviceability to a Product-Service System Model," Proc. 2013 Spring Servitization Conf. (SSC 13), 2013, pp. 95–100; www.aston-servitization.com/publication/file/35/31_spring-servitization-conference-2013-proceedings.pdf
8. P. Bosch-Sijtsema and J. Bosch, "User Involvement throughout the Innovation Process in High-Tech Industries," J. Product Innovation Management, vol. 32, no. 5, 2014, pp. 793–807.
9. L. Downes and P. Nunes, Big Bang Disruption: Strategy in the Age of Devastating Innovation, Penguin, 2014.
10. P.M. Bosch-Sijtsema and J. Bosch, "Plays Nice with Others? Multiple Ecosystems, Various Roles and Divergent Engagement Models," Technology Analysis and Strategic Management, vol. 27, no. 8, 2015, pp. 960–974.
11. D. Tilson, C. Sørensen, and K. Ly ytinen, "Change and Control Paradoxes in Mobile Infrastructure Innovation: The Android and iOS Mobile Operating Systems Cases," Proc. 45th Hawaii Int'l Conf. System Science (HICSS 12), 2012, pp. 1324–1333.
12. F. van der Linden et al., "Software Product Family Evaluation," Software Product-Family Engineering, LNCS 3014, Springer, 2004, pp. 110–129.
13. E. Backlund et al., "Automated User Interaction Analysis for Workfl ow- Based Web Portals," Software Business: Towards Continuous Value Delivery, Springer, 2014, pp. 148–162.
14. H.H. Olsson, H. Alahyari, and J. Bosch, "Climbing the 'Stairway to Heaven'—a Multiple-Case Study Exploring Barriers in the Transition from Agile Development towards Continuous Deployment of Software," Proc. 38th EUROMICRO Conf. Software Eng. and Advanced Applications (SEAA 12), 2012, pp. 392–399.

Part B:
Smart systems, autonomous vehicles, and robotics

Public safety and security

Gunnar Holmberg, Saab, gunnar.holmberg@saabgroup.com
Martin Rantzer, Saab, martin.rantzer@saabgroup.com
Martin Holmberg, RISE SICS East, martin.holmberg@ri.se

OVERVIEW

The area of public safety and security has been subject to attention for a long time. Many potential threats have become more severe and more frequent, while ambitions to avoid events and minimize consequences increase.

The wide variety of potential threats ranges from individual accidents via terrorist threats to large events and public gatherings to raging natural disasters. This stresses the need for being able to avoid and to respond to a vast range of possible situations. It also accentuates the need for rapid situation awareness, acting with scarce resources and continuously adapting to better situation awareness in order to achieve best possible response to an event.

Technology development offers new possibilities, and will continue to do so for the coming decades. Important trends include for instance better surveillance methods based on improved sensors, communication and drones, use of decision support and autonomy to improve understanding of situations and possible responses, and better predictability of potential adverse events. But in terms of security, technology may as well create new possibilities for adversaries.

An interesting development yet to be debated is the use of dedicated systems for emergency response versus the use of other systems that could play a key role. As an example, we could see systems ranging from two disjunct groups of systems, where the emergency response system tries to act with only ad-hoc interaction with for instance traffic (currently there might be a message on the radio to give priority for an ambulance), to a coordinated action (autonomous cars in a smart traffic system may coordinate to ensure free way for the ambulance), and finally where there are more active requirements on the systems depending on ability (for example if a passenger in an autonomous car gets a heart attack, the car drives automatically to a hospital and interacts with other traffic to obtain priority). In addition to this, there are other safety contributors, such as less human-induced accidents, but also potential to inhibit certain types of adverse events.

The use of new command-and-control systems for emergency response in the society will help first responders on all levels to get a better understanding of the situation in a larger area, thus making it possible to take into consideration more events, more possible future scenarios, more resources, and similar. However, this also means that a responder will have more systems and personnel to interact with, making the task more complex in many ways. It also raises the question how a command-and-control system should be designed in order to facilitate the interaction of diverse types of responders; can each organization have their own command-and-control system or are all parts of society too interlinked? The human-system interaction and the roles of the humans in the system will therefore have to be explored further.

A complete re-design of a new command-and-control system will be very costly, and we can thus anticipate that in the foreseeable future we will have to do with gradual improvements of the existing systems. One challenge will thus be to make sure that new developments will be along such lines that future interaction between systems will be possible.

Apart from the development of systems for safety and security, recent events have shown that cyber security can be as big a threat as physical events. These challenges will, however, not be part of this overview.

INTERNATIONAL MATURITY

The very definition of a crisis is that it is an event out of the ordinary with consequences so large that it is difficult to assess the problem and direct the response system. A number of new and emerging technologies have been used during crises as stand-alone solutions without incorporation into a mission-wide command-and-control system, such as data mining for analysis of social networks, drones for reconnaissance, and ad-hoc networks for communication. Experience has shown that these new technologies can help increasing the capabilities of the response system, but they are still not mature and robust enough to be deployed on a wider scale. Also, due to the lack of training and appropriate user interfaces, they are not easily used by users not experts in the specific technology.

Situation awareness and command-and-control systems have long been in the focus for military research and there are several examples of technically mature systems and products that use a systems-of-systems approach to solve distributed real-time challenges much like those within public safety and security. The solutions for public safety and security challenges are less mature. Possible reasons for this might be less funding, more disparate organizational structures or a more diverse set of possible disasters to handle.

Many nations around the world have high ambitions to handle disasters within their own borders but also offer international support. There is a large potential for coordination between the relevant actors between the countries, both when it comes to organization and integration of technology. It is also often a problem to receive support in an organized way when disasters strike and the infrastructure in a country or region is ruined.

CHALLENGES

New technologies are often developed separately in a bottom-up fashion. However, the very nature of the field of public safety and security requires that many new technologies are incorporated into a command-and-control system in order to give full benefit of their use. The co-development of several technologies represents a new challenge previously attempted mainly for military purposes or other applications where large sums of money have been allocated.

The robustness of the systems against the malfunction of infrastructure or other parts of the command-and-control system will be essential in a crisis (graceful degradation).

Cyber threats will be a possible show-stopper unless proper choices and countermeasures are proposed.

TYPICAL INHERENT TECHNOLOGIES

Situation awareness is a key area for public safety and security. A well-functioning human-computer interaction between users and a command-and-control system is key to achieve an understanding of the real-life situation and what resources are available to solve the problem. Improved situation awareness is dependent on research areas, such as mission planning to ensure a coherent high-level plan that influences underlying decisions by users or the system, intelligence and analytics using deep learning and data/information fusion to make sense of the vast amounts of historical and real-time data that need to be combined, visualized and interpreted. Possible tools to support this might be digital cognitive companions or augmented reality that can offer advice to the user and improve the understanding of what happens in a crisis area far away from a command-and-control center.

It is vital to maintain situation readiness for as many potential crisis scenarios as possible. This means that all resources that are usually needed for planning and execution will not be available due to, for instance, limited resources in general, limited communication bandwidth, potential effects of cyber-attacks, insufficient staffing, and similar. This means that the planning and prioritization must be made with the limited resources that are at hand, calling for the design of a system that is not dependent on all parts of the system, but rather will continue to give at least a limited functionality even when parts of the system fail or are not yet available. Also, even when the system has suffered a severe blow it must be resilient, meaning that i should allow for graceful degradation and have the capability to quickly regain as much functionality as possible after the shock to the system. Computer programmers have designed complex systems with at least some of these features for years, but in order for a complete crisis-management system to work this way, one has to consider also the training of the people, the organization, and the methods used in order to render planning with limited resources possible.

Intelligent autonomous systems must also be able to collaborate with other (autonomous) systems since there will not always be a human in the loop to make sure they avoid collisions and facilitate their task planning and cooperation. There is thus a need for accurate positioning of the systems and effective ways of communicating their respective position to other systems in the vicinity. Current autonomous systems rely heavily on GPS for positioning, with a possible fall-back to cellular wireless networks. For cases where these signals are not available, for example indoors, a number of methods can be used to localize the system in the surrounding based on sensor readings and a map that is either available beforehand or built as the system moves along. The sensors can be of one or several types (such as visual, radar, or lidar[1]), where sensor fusion of different independent sensor output can increase the accuracy. These methods are called SLAM (simultaneous localization and mapping), and can work fine for a limited time, but due to the exponential growth of errors in the calculation of the position, these methods will not work for a very long time unless they are given accurate reference points every now and then.

Even when the position can be determined with high precision, there is currently no universal traffic-management system for autonomous systems. It would be possible to avoid collisions by making individual platforms communicate with other platforms in the vicinity (de-centralized), or to have a centralized system which handles all platforms, or to have a solution with an architecture somewhere between these two extremes (for instance hierarchical). It is also possible that several traffic-management systems have to co-exist in order to avoid collisions on both short and long term, and also to make route planning possible on different time scales.

Autonomous systems operating in complex surroundings will have to be able to grasp the situation without relying on external support from a centralized command-and-control system in case the communication fails. This goes beyond the task of mapping the surroundings, but also includes the identification and tracking of objects in the vicinity. By understanding the situation and being able to make predictions on the movement of objects (humans, vehicles, and similar), the autonomous system will be able to plan its own movements and completing simple tasks such as following and monitoring objects of interest, identification of lost people, or finding areas affected by natural or man-made disasters. New technologies are available for many of these tasks, such as deep-learning neural networks, machine learning, and reinforcement learning.

Sensors are now distributed over large areas, and the utilization of many networked sensors offers new capabilities and precision in terms of identification, tracking, and situational awareness. As mentioned above, autonomous systems rely heavily on sensors, and will thus be distributing even more sensors into the environment. When dealing with a crisis, autonomous systems and their sensors sometimes are exposed to rougher environment than during everyday operation, calling for the development of more robust sensors. Networks of autonomous systems with advanced sensors can collaborate to perform complex tasks, if proper care is taken to provide communication and distributed-data-fusion methods. The design of a network requires consideration of demands of latency, bandwidth, total load, and availability (quality of service). The use of 5G networks for communication may help solving some of these issues, but further exploration of the limitations is necessary. Furthermore, the architectures of the communication network and of the distribution of the data fusion have to be chosen to properly manage the network. The network needs to be able to optimize the use of its resources, thus being able to plan the use of all sensors in the network, including the movement of the sensors, the operating mode of the sensors, the deployment of new sensors, and the choice of fusion methods.

The planning of tasks for autonomous systems is often made beforehand by a human operator, making the need for autonomous decisions during operation quite low and limited to simple decisions such as re-routing when an obstacle appears. In a more complex task where the operators and the autonomous systems have to handle a highly uncertain environment and also deal with the possibility of partial system failure, the autonomous systems have

to be able to make more advanced decisions to prioritize and plan tasks, possibly together with a human operator. Research on swarming of drones has shown that it is indeed possible for autonomous systems to interact and plan simple tasks even when the number of systems is great, but swarming has yet to be used for more complex tasks.

Human operators can easily interact with a single autonomous system to give tasks and interpret the feedback from the system. However, in the case where many autonomous systems have to be directed by a single operator, the design of the control interface is much more difficult. It requires the use of techniques such as mixed-initiative reasoning, where a combination of the user's ability to appropriately choose a task and the autonomous system's ability to know its limitations and capabilities is used to find a good balance between user needs and the capabilities of the autonomous system(s). The development of new methods for the interaction with, and control of, networks of autonomous systems is necessary when operations are made in large areas with many objects. The situation becomes even more complicated when there are several users in the same area using partially the same autonomous systems. Similar problems arise in many organizations where resources have to be shared, but the division of user rights is then made manually or by using predefined rules. In times of crisis when all tasks are time-critical, the scheduling and prioritization have to be made in real-time without interfering with or disturbing the work of the users.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

To really improve public safety and security would require breakthroughs in research areas such as intelligent collaborative autonomous systems with (swarms of) drones that could quickly provide improved situation awareness in a disaster area.

Big-data systems using perception and learning techniques to quickly asses a crisis situation, or even predict and prevent upcoming disasters, would also be an enabler. This would probably need to be combined into computer vision, identification, tracking, and localization of objects in complex scenarios.

A third prioritized research area would be resilient adaptive mission-critical cloud (managing very high variable load) in combination with wireless communication with good geographical coverage, low latency, high availability and bandwidth.

POTENTIAL FOR SWEDISH POSITIONING

Sweden has a long tradition in building large systems involving players from academy, society and industry (the triple-helix model), which is a requirement for finding technical solutions that span several fields of applications, technologies and users. The presence of Swedish industry and academic research in key areas such as communication, data fusion (including big data), sensor technologies, autonomous systems, and systems engineering provides a platform for the development of radically new solutions for future threats to public safety.

Sweden has the potential to create "society-wide" systems-of-systems that can manage available resources to achieve a "greater good" in crisis situations. This could range from prioritization of sensor and communication resources in the network to a coordinated action of autonomous cars in a smart traffic system that coordinate to ensure free way for an ambulance. Taking this one step further in the future when we do not own individual cars, any autonomous car may be "commandeered" as evacuation vehicle in a disaster area to bring injured to hospitals.

POTENTIAL PLAYERS

There are a number of key organizations in the public safety and security arena.

- Counties and municipalities have the regional and local responsibility to manage crises within their geographical areas.
- Fire brigades, police, hospitals and ambulance services are first responders to all crises.
- MSB (Swedish Civil Contingencies Agency) is responsible for issues concerning civil protection, public safety, emergency management and civil defence as long as no other authority has responsibility.
- Research on crisis management and technology are carried out by the Swedish Defence Research Agency (FOI), the Swedish Defence University (FHS), Research Institutes of Sweden (RISE) and Swedish universities.
- Saab Group is the major Swedish defence and security industry.
- Ericsson, Telia, Telenor and Tele2 are the major suppliers and operators of communications infrastructure in Sweden.
- Non-governmental organizations (NGOs) such as Missing People also perform important work within the field, constituting a type of end-users with potential input on research and development.

NEED FOR SUPPORTING ACTIVITIES

Public safety and security needs to become constantly more effective before, during and after any crisis. This includes research in diverse fields such as management and organization, social sciences, crisis medicine and relevant technical areas.

For the preparation for crises, there is a need for improved methods and tools for simulations, training, exercises, testbeds and early deployment of prototypes and new technology.

There is also a need for increased support for international disaster relief and our ability to support people in need both in our region and further away.

Footnotes and links:

1. Lidar is a surveying method that measures distance to a target by illuminating that target with a pulsed laser light, and measuring the reflected pulses with a sensor. en.wikipedia.org/wiki/Lidar

Smart cities

Stefan Thorburn, ABB, stefan.thorburn@se.abb.com

OVERVIEW

Urbanisation is a global trend where today more than 50% of the population live in cities, and the rate is constantly increasing. This results in new and interesting issues where new and improved solutions are needed. The type of solutions differs from traditional approaches in being multidisciplinary in nature. When implementing these solutions, cities sometimes start to brand themselves "smart" cities.

Smart cities are evolving toward highly interconnected infrastructures with an unprecedented level of human-machine cooperation where autonomous systems assume a key role. Smart cities are thus ultimate examples of cyber-physical-human systems, the dynamics of which is driven by controllable and uncontrollable external and internal forces.

Examples of such forces are human innovation, values, choices and behavior, collaborative and competitive socio-economic processes, climate change, extreme events and constraints due to resource depletion.

The ultimate operational goal for smart cities is to ensure a system level "intelligence", capable of adapting and balancing functional robustness with system fragility in a way that ensures long-term sustainability.

The transition toward smart cities will by necessity imply decades of increasingly complex co-existence and interaction between human and autonomous systems, which poses many new challenges and is poorly understood. The smart city will create new axes of interaction from traditional areas based on data mining – and should still acknowledge personal integrity.

Some cities which brand themselves as a "smart city" highlight key elements such as **high usage of digital technology, environmental sustainable operation** and **enhanced mobility.** The solutions should lead to simplified and improved lives for its residents, its visitors and businesses. With anticipated global warming effects, **resilience** is another key factor of increasing importance for the cities.

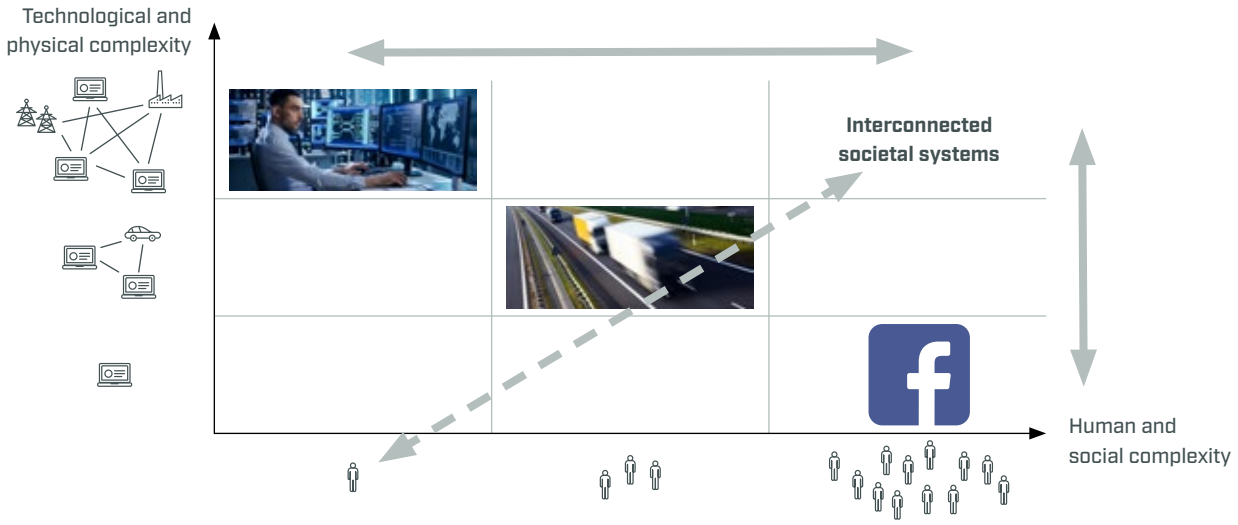


Figure 1: The smart city will mix human and social complexity with technological and physical complexity. Treating each axis individually can of course be problematic in itself, but the issue here is combining them. Also, the personal integrity of each individual should be maintained when deriving information for the "greater good".

INTERNATIONAL MATURITY

The concept of combining big data with cities has been around for long[1]. However, there are still few very good examples of systematic implementations. Driven by the better communication infrastructure, IoT, cloud management, mobile-phone apps and structural changes, several new smart city solutions are foreseen. Dubai[2] is one recent example where the goal is to develop a smart city where the building blocks are now developed. They are now rolling out an internet-of-things strategy and data-wealth initiative. To manage and make use of all these data flows, autonomous systems should play a key role organizing the information.

Interestingly, a key factor in the smart city may be the digital signatures and IDs needed for secure identification of persons and devices. Sweden and Estonia are two countries where the BankID system and the e-identity already plays a significant role in our daily life. Dubai are now introducing digital IDs and this is a global trend[3].

The "smart city" will also be a significant building block in the trend for more resource efficiency. Several parallel efforts evaluating different technologies are tried out in the EU-funded READY[4] project. It highlights several important aspects to increase resource efficiency in different demos mainly along the technological and physical complexity axis.

CHALLENGES

Some areas present typical challenges for the development of smart cities, such as:

Water – 40% of drinking water is needed on 1% of global surface

One particular example here is water-pipe leakage. More than 20% of the water in the city of Stockholm disappears in the pipes. To properly measure the 2,400 kilometres of pipes, to develop demand forecast with data mining, to minimize pumping energy and to quickly identify leakages will require advanced big-data analytics. Reduced leakage will save resources and also minimize emergency street repairs of broken pipes. Such repairs stops the traffic flow in the streets which leads into mobility constraints.

Mobility – the needed flow within the city

Smart sensors will control traffic lights but how shall they be optimized from a city perspective? Today, apps can suggest the best commuter options involving busses, subway, trains, cars, bikes and walking. How can we handle payments, travel optimization and contingencies like faulted transport service or a major concert in such autonomous system? Air pollution is the next major issue for the cities and diesel ban is now discussed or proposed in a number of major cities. How will the city manage transports without diesel? All these solutions must be developed under the existing infrastructure considering both public and private companies, communities, regional and national aspects.

Electrification – A drive to become 100% renewable city

The diesel fleet will probably be (partly) replaced with electric vehicles which are then connected to the emerging new electrical infrastructure. The cities are finding ways to utilize photovoltaic electric generation and storage inside the cities. Blockchain technology and micro payments are tested in several pilots. Electric-vehicle charging points are developed and this new infrastructure will be integrated and optimized with existing grid.

Demography changes – How IT will enable the future smart city accessible for all

In certain parts of the world, the population's mean age is increasing, changing the demography. It is suggested that IT and sensor networks are tools that will be needed for creating a safer and more accessible environment for elderly. More and more of the health care will also be performed at the residents' locations, moving the "healthcare to the homes of the patients". Here we will see the need for autonomous support helping the users with various tasks.

In all these examples, autonomy lies as a web connecting different disciplines in order to handle the smart city of the future.

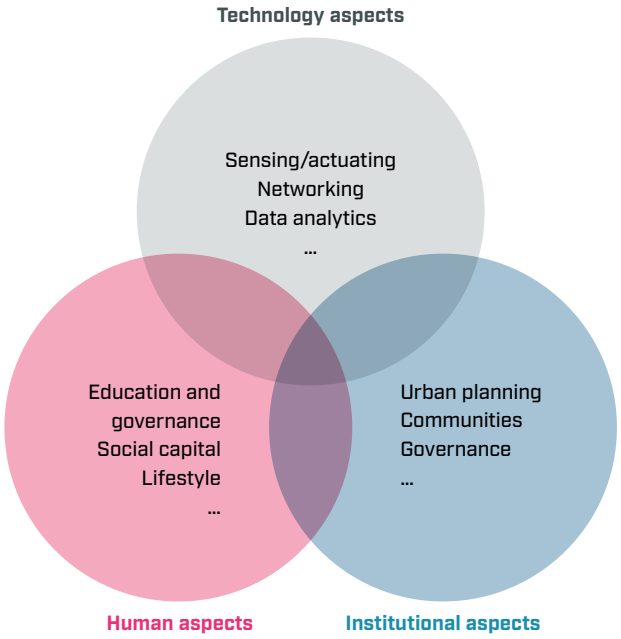


Figure 2: A smart city is not just a city that leverages new technologies, but is a complex ecosystem made up of many stakeholders including citizens, city authorities, local companies and industry and community groups.[5]

The need for a multidisciplinary approach

The biggest hurdle is most likely the interaction between large technical systems and human behaviour. This area requires an interdisciplinary approach. There are also a multitude of stakeholders and existing infrastructure which has to be included. This can cause hurdles when the benefactors and the beneficiaries belong to different stakeholder groups.

Common technical platforms and infrastructure standards should be important building blocks but due to the large number of stakeholders and inherent old technical systems progress may take time. Good pilots and examples to demonstrate benefits are therefore needed particularly around topics which include and enhance the existing infrastructure.

Privacy and non-skill issues

Data collection and data sharing are key building blocks. An important aspect is therefore data privacy versus the "value for all" of aggregated data. How do you make the best use of public data without violating personal integrity?

It is also important to make sure the smart city systems also include the non-IT skilled residents. Not all will have a full-fledged GPS-positioned phone actively feeding information into social media.

Trustworthiness and cyber security

Important factors for a successful implementation of a smart city will be our overall experience of the system. One such factor will be the trustworthiness and availability of the system. Autonomous systems will be needed to manage these aspects such as distributed denial of service attacks. For instance, low cyber security in IoT devices like web cameras have caused massive disturbance in the internet[6].

TYPICAL INHERENT TECHNOLOGIES

Connectivity

Different technologies like 5G, fiber networks and short-distance radio communication are needed everywhere. Note that the cyber-security efforts must start already at the edge nodes in order to maintain trustworthy systems.

Positioning

Many services will require means for positioning or localisation. Several options should be available.

Data ownership

Ownership of data might need further juridical, technical and social investigations.

Identities of devices, sensors and persons

To maintain trustworthiness of the systems and to handle aspects as personal integrity and financial transactions, the identitites of devices, sensors and persons must be known. This includes an "end-to-end" security approach to provide the ability to maintain and protect personal integrity. Proper identification will also be needed to grant access to information at different aggregation levels.

Integration of different systems

Different information sources, data types and varying ownerships must be able to interact. This must also work for old, existing infrastructure.

Cloud, edge and fog computing

Data analytics must be made in various location in the city environment. Here we need autonomous systems to evaluate which parameters that should be controlled and calculated locally and which parameters that need a global optimization.

Existing technology systems

A large number of existing technical systems will be integrated in the future smart city. These may be old, probably working beyond their intended life span and lack any type of connectivity. Examples are the electricity grid, water and wastewater, transports, buildings and district heating and cooling.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

The development of autonomous vehicles will somewhat be interlinked with the smart-city infrastructure. Development in each of them will be connected to the other. Examples of possible future stakeholders here are startup companies like Einride[7], different drone deliveries and pizza home deliveries. Such services will most likely interact with the smart city.

Machine learning and data analytics emphasizing dynamic system aspects are also needed. Uncertainty and inhomogeneous data availability influence the analysis.

Different systems are coupled within the city. In energy management, for example, data from other infrastructures will enable better decision-making and control of electricity supply, and vice-versa. Such couplings should be investigated and further developed.

Cyber security and robustness/trustworthiness will be key factors. With the increasing number of connected actuators, the exposed surface increases significantly. Note especially the connection between different cyber-physical systems and how one system can disturb the other.

Blockchain technologies may be a platform for many different services in the smart city. It will then be an important tool in transactions, sharing services and traceability.

Biometric identification is already under development and may be an important technology in the smart city.

POTENTIAL FOR SWEDISH POSITIONING

In Sweden there is a strategic innovation program around smart and sustainable cities called ViableCities[8]. The program has recently been started.

Society critical infrastructure like water systems, electricity, district heating, civil emergency protection and communication infrastructure are intertwined. A disturbance in one system may propagate into the others. By applying methods from the industry control rooms, cities can have a much better overview and hence increase resilience. Vinnova is therefore running a pilot for a "city control room" in the city of Västerås[9].

Health and elderly care is another area where Sweden may be another potential area for positioning. Sweden spends a relatively large share of public money on the elderly care, and the IT maturity is also relatively high in Sweden. The combination of smart city technology, public funding possibilities and relative highly IT-matured people can be a game changer. Startup companies like Aifloo[10] show one way forward.

POTENTIAL PLAYERS

Digital Demo Stockholm[11]

Digital Demo Stockholm is an example of collaboration between city/region, industrial partners and academia to solve societal challenges through digital technology. Present focus is on water, mobility, social integration, safety for elderly, energy efficiency and digital healthcare.

C40 Cities[12]

C40 is a network of the world's megacities committed to addressing climate change. C40 supports cities to collaborate effectively, share knowledge and drive meaningful, measurable and sustainable action on climate change. Research, data and similar are exchanged and shared.

DataCity[13]

DataCity is a global open innovation program that brings together cities, companies and startups to address together city challenges and develop solutions to build sustainable and efficient cities, using data and technologies.

NEED FOR SUPPORTING ACTIVITIES

The text above reflects mostly the author's technical viewpoint and lacks a bit of the social aspects of a smart city. The author also believes that the knowledge from human and social complexity lacks some of the technological skills to understand and design the tools and hardware for a smart city. Hence there is a need for society to mix these two aspects and see how it can develop.

Around each topic is it possible to develop a dedicated structure of supporting activities. However, as an example, let us view a few questions arising around the patient that is treated in his/hers home, with the corresponding challenge areas in parentheses:

- How shall the user experience be designed such that an elderly or disabled person will be able to manage the system (technology solutions and usability interface)?
- How shall the health care handle large amount of external data flowing in (autonomous systems and legal structure, medical education)?
- How shall the legacy home care and health care organizations coexist (social scientists, governmental structure, organizational coaches, ...)?
- How to handle personal integrity and cyber security (technology, behavior, trust)?
- How shall we build proof of concept when the financial split of benefits and costs will change from today (external funding need)?
- How do we prove that the solution is beneficial for society (big-data analytics, social scientists and economists)?

Footnotes and links:

1. blogs.oii.ox.ac.uk/policy/promises-threats-big-data-for-public-policy-making
2. www.metering.com/news/dubai-iot-strategy
3. www.gemalto.com/govt/identity/digital-identity-trends
4. www.smartcity-ready.eu/about-ready
5. www.ieee.org/publications_standards/publications/periodicals/ieee-smart-cities-trend-paper-2017.pdf
6. www.wired.com/2016/12/botnet-broke-internet-isnt-going-away
7. www.einride.eu
8. www.viablecities.com
9. www.sics.se/projects/stadens-kontrollrum
10. www.aifloo.com
11. www.digitaldemostockholm.se
12. www.c40.org
13. datacity.numa.co

Autonomous vehicles – the vehicle perspective

Karl H Johansson, KTH, karl.henrik.johansson@ee.kth.se

DEFINITIONS

Autonomous vehicle: A vehicle driving by itself.

Automated vehicle: A vehicle in which certain vehicle functions are automatically controlled or operated.

OVERVIEW

An autonomous vehicle is a vehicle that can move by itself in unstructured environments without continuous human guidance. Such capability is desirable for all type of vehicles transporting people or goods: land vehicles, watercraft, aircraft and similar. The main motivation for autonomy is often higher convenience, increased safety, better fuel efficiency, and reduced labor cost.

There is an important distinction to make between autonomous and automated vehicles. Autonomous means that the vehicle is driving by itself, while automated means that certain vehicle functions are automatically controlled or operated. Most of today's relevant projects in the automotive domain are dealing with various levels of automation and how such automation can benefit from real-time information exchange between vehicles and between vehicles and the infrastructure. Strictly speaking, it is seldom making vehicles more independent that puts the technological advancement on a higher level of autonomy; this is done by a higher level of connectivity together with decision-making based on more complex data. Thanks to the fact that the vehicle is considered as part of a transport system, it will be possible to improve the vehicle's ability as well as the performance of the overall system.

The Society of Automotive Engineers (SAE) has introduced an automated-vehicle classification scheme with six levels (see figure 1). Level 0 corresponds to the situation where the system may issue warnings but has no vehicle control, while Level 5 corresponds to the situation where no human intervention is required. Levels 2–4 are sometimes referred to as "hands off – eyes off – mind off" due to the decreasing amount of driver involvement in these levels.

INTERNATIONAL MATURITY

Depending on vehicle types, the technical maturity of automation is drastically different: the first aircraft autopilot was developed more than a century ago and many passenger aircraft are highly automated since decades; fully automated shuttle buses are commonly available for inter-terminal transport at large airports; but there is no commercially available car, bus, or truck that is completely automated and able to handle all normal driving modes (corresponding to SAE Level 5).

There are three major categories of leading players in the industrial area of autonomous vehicles: sector-specific companies (automotive companies, defence companies and similar), IT industry (Google, Apple and similar), and start-up companies. Most of these companies have close ties to leading academic research environments. For example, several start-ups have been formed out of university teams who participated in the DARPA Grand Challenge on autonomous vehicles[2] and some of these start-ups have later been acquired by larger companies such as Google, Uber, and Ford.

From an international perspective, US leads this development with a particularly agile IT industry and innovation hubs in the San Francisco Bay Area and elsewhere, but there are major developments also around the automotive industry in Europe and Asia. Some cities, like Singapore and London, are taking particularly strong moves in the adoption of "smart city" technology, where traffic management and automated transport are key components.

An important reason for more intelligent transport solutions for any large city or country is the financial, industrial and societal importance; as an example, goods transport in the EU amounts to 3.5 trillion tonne-kilometres per year with three million people employed, whereas people transport amounts to 6.5 trillion passenger-km with two million employees.[3]

CHALLENGES

To understand the challenges and obstacles in the development of autonomous vehicles and in general automated transport systems, it is important to understand how various underlying technology areas are connected and how the realization of the broad vision depends on the development of individual components as well as on how well they can be integrated in the overall system. Some of these aspects are illustrated in the diagram in figure 2.[4] Note that an automated highway system[5] or a cooperative road-freight transport system[6] reside in the very centre of this diagram, intersecting the areas of automated vehicle systems, connected vehicles systems, and intelligent transport systems.

In each of these areas, there are major obstacles to overcome.

- The most important one has to do with traffic **safety:** how to guarantee safety under extremely uncertain external conditions (weather, road, traffic etc) for a software-intensive system of enormous physical, dynamic, and human complexity. It is impossible for system designers to anticipate all possible scenarios at the design or testing phase, and current design methods do not handle such large uncertainty and complexity.
- A related challenge is on **reliability:** how to build an automated system with similar (or even higher) level of reliability than we have with manually driven vehicles. Today's cars run 3.4 million vehicle hours between fatal crashes (corresponding to 390 years of non-stop driving) and 61 400 vehicle hours between injury crashes (seven years of non-stop driving).[7]
- Another essential obstacle to overcome is on **security,** privacy, and trust; for instance how to handle adversarial attacks on vehicle-to-vehicle communication systems.
- A fourth potential obstacle is on how to integrate the

SAE level	Name	Narrative definition	Execution of steering and acceleration/ deceleration	*Monitoring* of driving environment	Fallback performance of *dynamic driving task*	System capability *(driving modes)*
***Human driver* monitors the driving environment**						
0	**No automation**	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	**Driver assistance**	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes
2	**Partial automation**	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes
***Automated driving system* ("system") monitors the driving environment**						
3	**Conditional automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the dynamic driving task with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes
4	**High automation**	the *driving mode*-specific performance by an automated driving system of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes
5	**Full automation**	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**

Figure 1: The Society of Automotive Engineers' (SAE) six-level automated-vehicle classification scheme.[1]

human interaction into such a system; for instance how to safely move from different levels of automation, and how to interact between drivers and other operators at these mode shifts.

There are related but more general challenges related to the human involvement in these systems, such as the trust and acceptance of the new technology as well as the societal aspects related to the potential convergence between public and private transport modalities.

Some of these challenges are so hard that many researchers believe that it will take decades until we will see fully automated vehicles in the transport system.

TYPICAL INHERENT TECHNOLOGIES

There are both basic and applied research needs to target in order to address the key challenges above. The four research areas discussed below are all on relatively low TRL levels but advances in these areas would have significant impact.

High-confidence cyber-physical infrastructure systems

The automated transport system belongs to a class of so called cyber-physical systems that more and more come to operate our infrastructures, with tight integration between physical flows and coordinating cyber network. There is a major and urgent need to research the fundamentals of

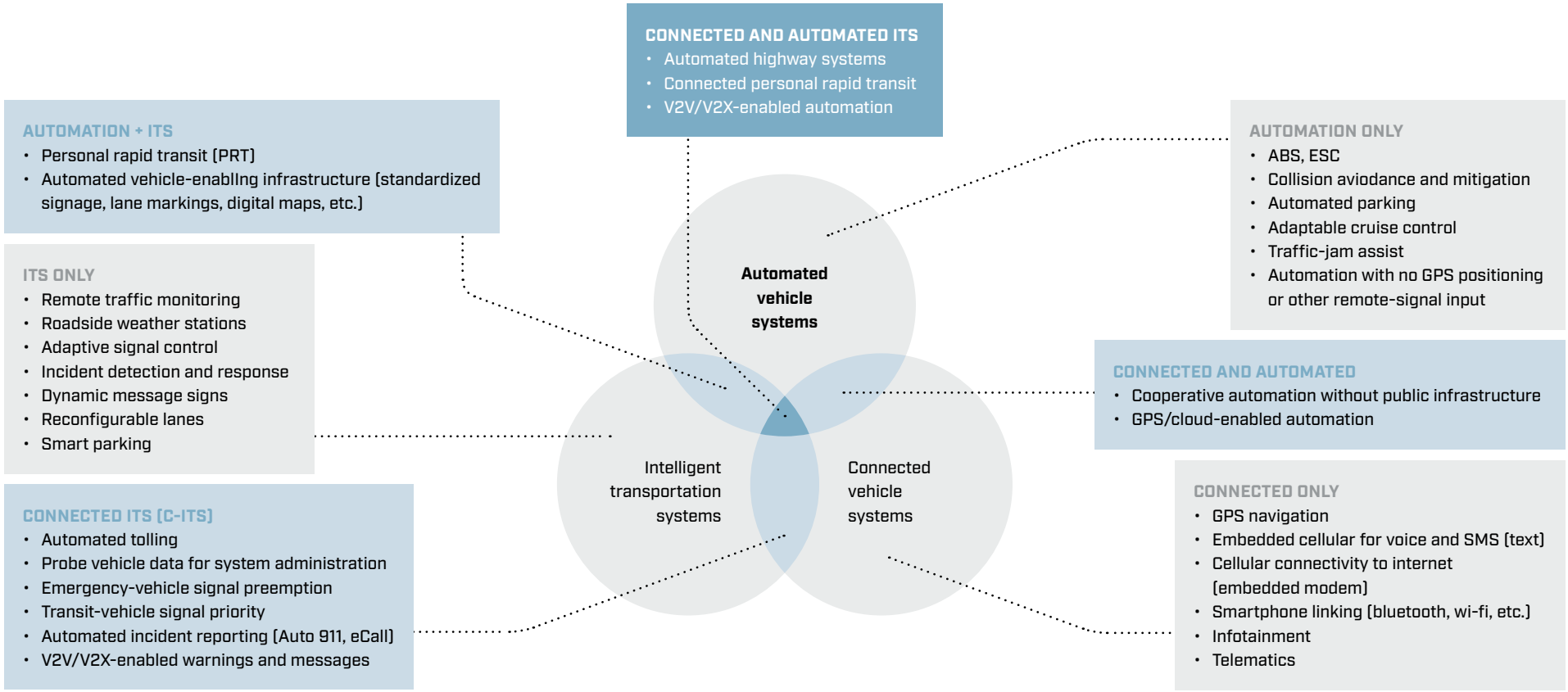


Figure 2: It is important to understand how various underlying technology areas are connected and how the realization of the broad vision depends on the development of individual components as well as on how well they can be integrated in the overall system.

how to design high-confidence cyber-physical infrastructure systems.

Today's systems and computer sciences do not provide the right tools for handling these multi-disciplinary systems of very large scale generating an enormous amount of data. Because of their societal importance and safety-critical nature, these systems need to act in a provably correct manner and should be able to handle faults and other degradations in a predictable way. There is a need for new model-based software development methods for design, verification and validation to overcome today's limitations of formal methods and brute-force and time-consuming testing.

Cyber-physical infrastructure systems work over multiple timescales and large geographic areas, so the supporting information and communication technology needs to be flexible in giving real-time performance or resource-constrained computing whenever needed. For example, today's vehicle-to-vehicle communication standards are not reliable enough or scalable to handle future complex automated transport systems.[8]

Human-machine interactions with safety and performance guarantees

The design of human-machine interfaces is an active research area since many years, partially due to the plethora of new internet-connected gadgets and devices all around us. Much of the development, however, has focused on best-effort technology and much less is known how to systematically handle human-machine interactions with safety and performance guarantees, which is needed in many foreseen traffic scenarios with multiple vehicles both with and without drivers.

It is widely debated if vehicles on different SAE levels will be able to safely co-exist in the transport system, as the multi-human interactions in such system become enormously complex and hard to predict. In addition, with the increased penetration of GPS-enabled smartphones and vehicles, traffic information and routing applications have become ubiquitous and allow drivers to individually optimize their driving decisions. The consequence of this evolution on a societal scale is much less understood as is how to incentivize drivers and traffic flows for global performance and resilience.

Reliable and safe data-based learning for real-time control

Automated and connected vehicles will depend and make decisions based on variety of data, from slowly varying map databases to real-time sensor measurements. Many of these data sources will be processed with machine learning algorithms.

Despite recent success of artificial intelligence and learning algorithms in areas such as computer vision and speech processing, there is little known yet how to utilize these techniques for reliable and safe data-based learning for real-time control. The closed-loop dynamics add dramatically to the complexity of the problem and might even generate completely new types of behaviours. Existing black-box learning techniques provide often no insight into the particular model or decision logic that has been obtained. For control of physical systems such insights are necessary to develop appropriate system robustness as well as fault-detection and diagnosis algorithms.

Cyber-secure large-scale automated systems

Cyber security is obviously an essential property of any computer system. Cyber-secure large-scale automated systems is particularly challenging because the attack space spans both the cyber and the physical world, and an attack might affect the physical infrastructure with devastating consequences.

Even if individual communication channels or specific devices are secure, this does not guarantee that the overall system is secure. As no system of sufficient scale and complexity can be made completely secure, it is important to have tools to reason about where to make security investments and how to cost-efficiently counteract adversaries.

A related problem is how to operate such infrastructure under privacy constraints on user data. Today, vehicles upload driving plans and state information openly to navigation applications, but as the amount of data grows, privacy considerations need to be incorporated.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

It is essential with a rather broad research agenda in road vehicle automation, as the current development is very fast and with a significant amount of speculations around certain technologies yet to be proven suitable. Recent research seems to suggest that some technologies are more viable and ready than others. When it comes to technology on a higher TRL level, closer to applications and prototypes, a few specific enablers are evident as discussed next.

Connected vehicle capabilities

Connected vehicle capabilities through vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication technologies seem to be close to providing the backbone for vehicle cooperation and intelligent transport systems. Further research is needed, however, on how to make V2V, initially focused on broadcasting traffic incidents and other alarms, capable of handling real-time safety-critical vehicle control. Current standards are not able to handle complex traffic scenarios with many vehicles and conflicting objectives.

Targeted automation technologies

Research is also needed to take V2I from rather slow monitoring and data gathering to agile decision-making in complex traffic scenarios. Targeted automation technologies for vehicle operations with professional drivers and direct economic benefits seem to be close to adoption. These technologies include autonomous truck loaders in restricted areas and highway truck platooning.

Managed and dedicated lanes

Related early public benefits from vehicle automation are in the development of managed and dedicated lanes, where equipped vehicles could concentrate together to benefit from an environment free from human drivers.

Cyber-physical systems, human-machine interaction and learning

For SAE Level 5 full automation, breakthroughs are needed in the technologies on lower TRL levels discussed in previous section. They are in:

- high-confidence cyber-physical infrastructure systems, where model-based tools for software development, verification, testing, and operation are needed to guarantee safety;
- human-machine interactions with safety and performance guarantees, allowing autonomous and human-driven vehicles to co-exist;
- reliable and safe data-based learning for real-time control handling the challenges of validating inductive learning and avoiding costly data labelling;
- cyber-secure large-scale automated systems guaranteeing proper operations and mitigation strategies under adversary scenarios, but also how to operate informa-

tion-rich infrastructures under privacy constraints.

POTENTIAL FOR SWEDISH POSITIONING

Sweden has the potential to leverage excellent scientific and industrial standing in several technology and research area needed for autonomous and connected vehicles.

Particular unique opportunities exist in areas of system development and integration. Swedish automotive industry is in a strong position to further develop its standing in automated vehicles in general. The integration of vehicles into intelligent transport systems suggests that tight interaction with infrastructure systems is needed. Such a development should benefit from that the Swedish infrastructures are in general in good conditions and thus would allow for the development, testing, and early adoption of advanced technologies. Similar to historical developments of the power grid and the mobile communication system, Sweden could have an international competitive advantage of industry and authorities working tightly together with leading academics.

In the creation of the future intelligent transport system, it is clear that 5G can play a crucial role, but that the development of such systems needs significant cross-disciplinary competences cutting across EE, CS, ME, CE and behavioural, economic, and social sciences.

Sweden needs to position itself in the area of development and rapid prototyping of services and applications for the emerging transport system. Particularly, those future applications with tight system integration needing a crossdomain approach for the developments should be particularly suitable.

Much of the success in the area of automated and connected vehicles (as well as other IT-intensive areas) found today in Silicon Valley, Singapore and similar, seems to follow a focus on broadly building internationally attractive academic conditions, the recruitment of world talent, and open innovative environments. There is a window of opportunity for creating such an environment in Sweden in this area, positioned to educate leaders of tomorrow, leveraging Sweden's strong traditions in the underlying fields and industries.

POTENTIAL PLAYERS

Sweden seems to be well prepared for future positioning in this area, but there is a need for looking across industrial domains and academic disciplines. Some alliances exist already through initiative such as Drive Sweden and VINNOVA's FFI, but currently they do not provide any major funding for blue-sky research and development in the area. As cross-sectorial initiatives are needed to target key technological challenges, there must also be support for such initiatives coming from various agencies and organizations.

NEED FOR SUPPORTING ACTIVITIES

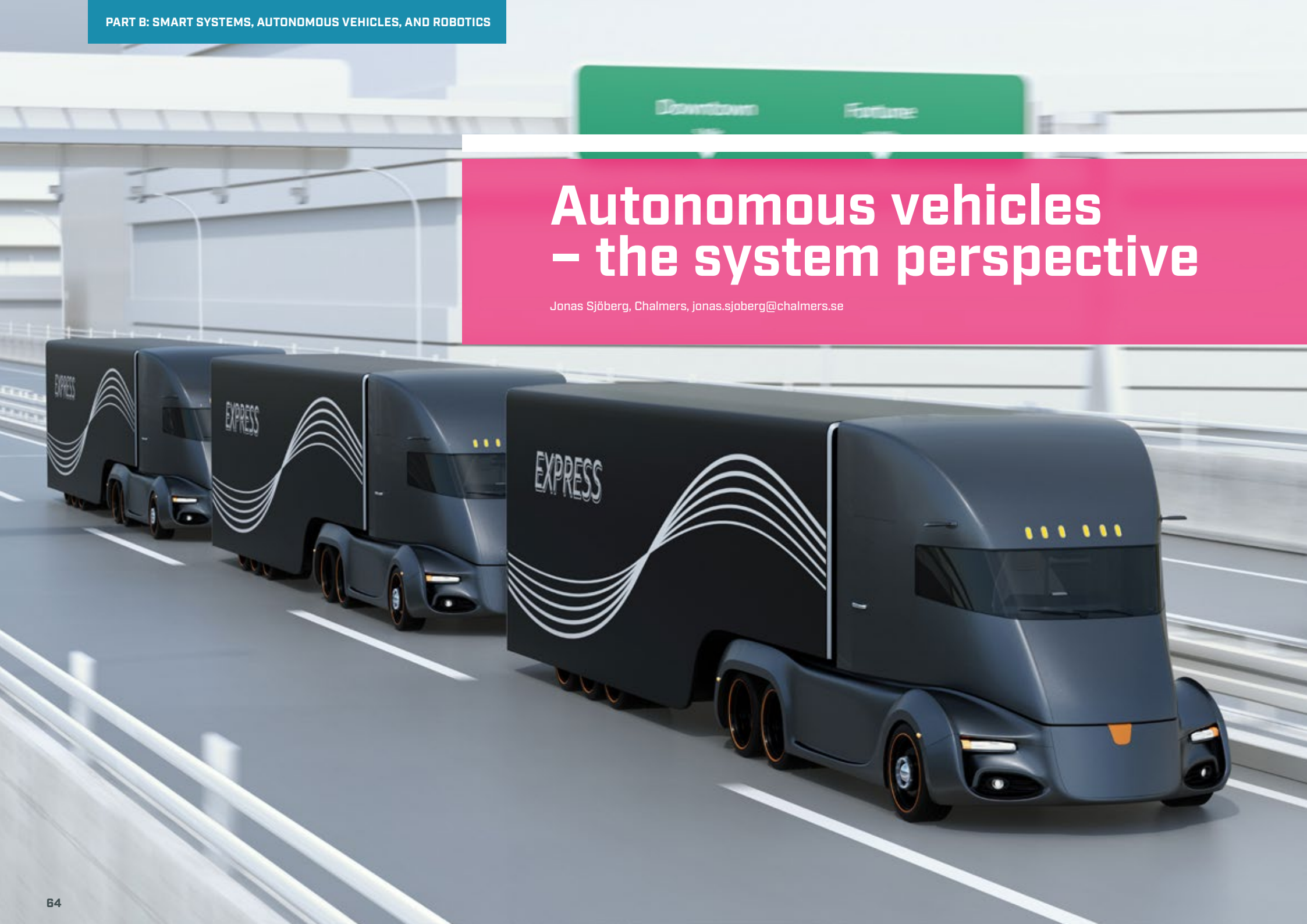
Given the broad spectrum of technologies and research areas related to autonomous vehicles with a quite diverse readiness level, there is a need to both build strong basic-research environments based in the core underlying disciplines focused on key societal challenges, and testbeds and demo sites to evaluate results and allow tight collaborations with industry and society. Current such initiatives in Sweden have a tendency to be too small and not at a level of ambition compared to what can be found in continental Europe, Southeast Asia, or USA. Additionally, a willingness to take certain amount of risks in such research initiative is required, by allowing the study of preliminary ideas and concepts.

Footnotes and links:

1. SAE J3016, SAE International, 2014.
2. S. Thrun, "Toward robotic cars", Commun. ACM, 53:4, 99—106, 2010.
3. European Commission, EU Transportin Figures—Statistical Pocketbook, Publications Office of the European Union, Luxembourg, 2014.
4. Connected Auto, Berkeley, 2016.
5. P. Varaiya, "Smart cars on smart roads: problems of control," IEEE Trans. Autom. Control, 38(2), 195–207, 1993
6. B. Besselink, V. Turri, S.H. van de Hoef, K.-Y. Liang, A. Alam, J. Mårtensson, and K. H. Johansson, Cyberphysical control of road freight transport. Proceedings of IEEE, 104:5, 1128-1141, 2016.
7. S. E. Shladover, "Road vehicle automation: history, opportunities and challenges", DREAMS Seminar,
8. Berkeley, 2017. K.-Y. Liang, S.H. van de Hoef, H. Terelius, V. Turri, B. Besselink, J. Mårtensson, and K. H. Johansson, "Networked control challenges in collaborative road freight transport." European Journal of Control, 30, 2-14, 2016.

Autonomous vehicles – the system perspective

Jonas Sjöberg, Chalmers, jonas.sjoberg@chalmers.se



DEFINITIONS

Platooning: A number of road vehicles following each other at close distance. All, except maybe the first vehicle, are driven automatically so that the close distance does not compromise safety. The motivation to forming platoons is to save fuel by lowering the air drag, and to increase the traffic efficiency by using less space. The automatic driving of the vehicles relies on sensor information and communication with the other vehicles.

OVERVIEW

The most challenging environment for autonomous driving is in mixed traffic where you also have human-driven vehicles, and maybe also pedestrians and bikers. The reason for this is that the autonomous vehicle needs to plan its future trajectory, and, for safety, predictions of all surrounding objects are needed, which are hard to obtain in a mixed environment.

From this general insight, one can naturally divide the effort towards autonomous driving into different setups.

1. Automised traffic in setups that are easy to predict. This can be, for example, robots in a warehouse so that the environment is very well known and non-predictable objects, such as humans, are absent. Such systems already exist today.
2. Less predictable setups, but still highly regulated areas, such as a harbor, a work-yard, or a mine. For these examples, low-speed solutions are often acceptable, which makes the predictions less critical. Instead these setups typically require planning and cooperation of actions involving several vehicles to, for example, avoid deadlocks.
3. Mixed traffic setups where there are good possibilities to predict human-controlled objects. This typically means highways so that bikers and pedestrians can be excluded, and the behaviour of the vehicles is somewhat restricted (for examples they all drive in the same direction) so that predictions can be done.
4. Last we have the most challenging setup with all kind of mixed traffic.

From these four setups one can easily identify strategies to facilitate autonomous driving:

- Change the traffic infrastructure so that it becomes easier to predict other objects. This includes possible changes that are costly and, hence, not attractive of that reason. But our traffic system has been changed a lot historically to separate different kinds of traffic to enhance safety and efficiency so such changes are not excluded.
- Improve sensors and software algorithms so that the awareness of the traffic situation increases. This also includes improvements of the modelling of all kinds of traffic objects so that the prediction quality improves. For example, by a more detailed modelling of pedestrian and bikers, one can potentially predict a turn earlier, which gives better predictions.
- Include vehicle-to-everything communication (v2x).
- Better decision algorithms that take more information and uncertainty into account.

Most research towards autonomous driving can be categorized into one of these four strategies.

INTERNATIONAL MATURITY

The development for autonomous driving is extremely fast. A description of the state-of-the-art is outdated as soon as it's completed. There are new experimental activities on autonomous driving announced almost every week.

So far, one speaks mostly about individual vehicles and how these interact with neighboring (mostly) human-driven vehicles. In this scenario, the challenge for the autonomous vehicle is to interpret the behaviour of human drivers, and to behave in a way that human drivers understand. There are many research projects considering the decision-taking of autonomous vehicles so that it will become more human-like.

The next evolutionary step is when vehicles are allowed to communicate. The intentions of other vehicles will be known and, hence, the predictions of their future trajectories will be much more certain. From this, the possibilities increase enormously. Uncertainty of all kinds decreases so that performance can increase. The road network can be better used since safety margins can be lowered. The performance of traffic system can be increased in all kind of ways.

There many research projects considering communication and there are several aspects that need to be investigated properly. At a first level, communication can be seen as additional sensors giving information from other vehicles and/or the environment. The second step is when the decision-algorithms make use of the communication that is called collaborative decision-making, which means that several vehicles collaborate to make decisions on how to act. For example, such algorithms have been investigated a lot for efficient and safe crossings, where each vehicles would like to deviate from it preferred speed as little as possible, but at the same time, traverse the crossing safely.

The communication is never 100% reliable, and the effects on the predictions and the system performance due to this is investigated in many research projects.

Also security aspects are being investigated, since when decisions are based on sent information, the trustfulness of the sender is crucial.

A specific topic using communication is platooning, which has been investigated and tested already several decades ago, but it is still not on the market. It is commercially mainly interesting for long-haul trucks, and it is likely that there will be commercial products on the market rather soon. All activities are on the research level, but some of them are mature enough for test-driving in real traffic to be performed.

CHALLENGES

One challenge is to create access to areas or roads where autonomous driving is allowed, and where there are benefits to me made from autonomous driving. This area should be prepared in such a way that the autonomous driving is possible without risking accidents including humans. Highways already qualifies into this class; further, one could have roads with bi-directional traffic prepared for autonomous vehicles.

If there is a clear business case, activities will start and the development will benefit from the experience from this first commercial use. To create business cases, authorities and stakeholders need to cooperate so that infrastructure and traffic rules are adapted. One can probably come up with many possibilities, but one, fairly simple, setting would be to identify a route between two points where there is a large demand of taxi service, and, additionally, there can be defined a lane or a road which can be reserved for automatically driven vehicles.

Generally, the area of autonomous driving is in the need of standards defining what an autonomous vehicle is allowed to do. With such standards, it will be easier for both humans and other autonomous vehicles to predict how an autonomous vehicle may act in a traffic situation. This, again, is with the goal that quality of predictions can be improved.

TYPICAL INHERENT TECHNOLOGIES

- Communication between vehicles and/or infrastructure. This communication needs to be fail-safe and secure, or rather, the quality of the communication needs to be known, so that the decision-making can adapt to the safety margins to the uncertainty and still assure safety.
- Control algorithms using the communication, scalable and optimized together with the communication for system performance.
- Standard for communication and when and what to communicate.
- Determination of which information should be shared by a vehicle. What are the risks of sharing?
- Aspects on information sharing. Traffic system aspects to optimize the use of the traffic system.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

Most of the technology is already available, but the cost is too high, and reliability is too low. Technical improvements are needed, to make the solutions more attractive. As described above, any progress that improves the predictions of the traffic situation is of great importance.

POTENTIAL FOR SWEDISH POSITIONING

There will be data and algorithms in individual vehicles, and there will be data and algorithms on higher level concerning a larger area. It is an open question which data should stay local and which should be shared, and in the same manner it is not clear which algorithms should run locally in the vehicles, and which of them should run outside the vehicles, maybe in the “cloud”.

One possible position for Swedish industry is to develop the information system, the infrastructure for this division of data and algorithm. Maybe it would be possible to take initiative for defining standards on this, to gain advantages, similar to those obtained by Ericsson and Nokia when leading the development of the GSM system for cellular telephones.

POTENTIAL PLAYERS

The WASP-connected companies are well qualified players, and of course, our universities: a new generation of engineers mastering the area is needed. Maybe one can stimulate start-up companies; there are many possibilities we have not realized yet, and sometimes these fall outside the interests of the existing companies. Saab and Ericsson in cooperation with the vehicle industry should have a good potential for success.

NEED FOR SUPPORTING ACTIVITIES

Cameras are general-purpose sensors heavily used for autonomous vehicles. Of course, recorded camera data is then important for the research. Also, for back-tracking, to investigate the reason for an accident, it may be of great importance to evaluate which data were used to develop the algorithm involved in an accident. Hence, it is important that camera data can be stored for future use. This is in conflict with the need of privacy that limits the possibilities to store camera images. A solution to this conflict must be found.

Autonomous shipping

Kalevi Tervo, ABB, kalevi.tervo@fi.abb.com

OVERVIEW

Autonomous shipping, or in a wider context, autonomous marine operations, includes development of higher-level automation and intelligence for shipping. The area is not only focusing on a single ship (which is important as well) and trying to automate that, but rather looking in a wider context of automation/autonomization of fleets of ships and extending that towards automation of logistic chains from supplier to producer.

The benefits of increasing the level of automation in shipping, in general, are decreased operation costs, increased safety, increased availability and more optimized value chains. Ultimately, when talking about autonomous and unmanned vessels and fleets (cargo shipping), the ships can be built without the facilities for people which decreases the cost of building the ship and increases the cargo capacity.

INTERNATIONAL MATURITY

The autonomous shipping development is continuation of the trend of increasing the level of automation and digitalization in shipping that has been ongoing for more than ten years. There are already advanced radars and electronic-chart display systems (ECDIS) solutions which can warn the captain of a possible collision so that action can be taken accordingly. There have been autopilots and dynamic positioning systems for decades already which can operate the ship based on references of course, speed, heading, or even operate the ship along a track. There are Eo-class notations for "unattended machinery spaces" which define the system and condition on building machinery spaces which do not need continuous human presence.

Apart from this, the international maturity of autonomous shipping is not very high – see challenges below. There are pilot projects such as US Navy Seahunter, Norwegian Yara Birkeland and similar which demonstrate the possibilities of autonomous shipping. When it comes to controlling along a track on normal situations in a good weather, the technological readiness is high. However, when it comes to replacing captain's decision-making process with an algorithm, it is not very simple task. There are attempts to do that but general solution do not exist. In addition, the reliability of the systems on board as well as the navigation and situation-awareness solutions are not high enough for autonomous shipping yet. In addition, the major challenge is the connectivity from ship to shore. When at open sea, the connectivity and automatic navigation and positioning rely on satellite technology which is not reliable enough at the moment, especially when going to the polar regions.

CHALLENGES

When it comes to increasing the level of automation in navigation, the most challenging task is to replicate or replace the bridge-personnel watch-keeping duties. The captain (or an officer of the watch) is responsible to continuously look out the windows, monitor radars, maps, and similar to make sure that nothing has to be done. Or in case if action is needed, he/she will perform the action. Open sea is one of the most violent environments imaginable and that creates significant challenges in building reliable situation-awareness systems to create a realistic picture of what is happening around the ship. The radars are pretty good at seeing long distances, but they do not have good capability to monitor close to the ship. In addition, the ability of the radar to recognize for instance a sailing boat depends on the radar settings, weather, and other conditions. One of the tasks of the captain is to use his/her experience to adjust the radar based on the operating conditions. For close-range monitoring there are several studies and pilots being made in using sensor systems similar to those having been used in automotive industry.

One of the major challenges when it comes to development of autonomous ships is the missing international rules and regulations. Therefore, it is seen that the first applications will be inland waterway transportation or coastal operations which are under flag-state rules and therefore more likely to allow exceptions. International traffic is under IMO rules and it will take 5–10 years to have the regulations in place.

Small ships operating in coastal areas can already today be remotely controlled. There are several companies offering remote-control solutions with some assisted autonomous functionalities even today. The challenge is to have a general and scalable solution that works in big and small ships. Another challenge, especially at the open sea is to make the availability and reliability of the systems higher. Basically that means developing system architectures to withstand more failures. The best way to do that is to utilize electric solutions as much as possible. In addition, predictive maintenance capabilities need to be improved in order to be able to prevent failures at sea. The systems onboard need to be designed so that the availability of the ship can be guaranteed even in case of a failure. In order to do this, in addition to design philosophy, rules and regulations of how the systems are built and also technological development of systems including power train, propulsion, automation and control systems are required for reaching higher availability of the ships. The systems should be redundant and modular so that any failed component can be swapped to a new one as "plug-and-play" during one port call. In addition, remote or automatic maintenance of systems should be developed in order to maximize the operability of critical components. In the end, if the systems still fail in the middle of the ocean, there should be a way to recover automatically and start the operation again. And in the worst case, there should be safe shutdown and escape functionalities to minimize the damage if all fall-back functionalities fail.

The navigation (finding the route, operating along it, avoiding collisions, and similar) can, in principle, be automated. However, in shipping the fundamental principle is that you should not base any critical operation on a single technology only. Therefore, the navigation cannot be based only on Global Navigation Satellite System (GNSS). Moreover, the connectivity to the ships cannot be based on satellite connectivity only. In addition, there are cyber-security issues which need to be taken into account. The development will most likely go first towards increasing

the remote-operation capabilities and there would be an "onshore captain" who is responsible of the ship (or a fleet of ships). That requires development of reliable and redundant connectivity solutions and technologies to provide the onshore captain a good understanding of the situation onboard the vessel.

When the ships operate in crowded areas, close to archipelago or are close to ports, there is significantly more traffic and also the allowed passages and operation areas are tighter. Moreover, the allowed passage for the ship depends on the ship's size and draft. The operation on crowded areas, coastal areas and close to ports is high-activity operation for the bridge and also for other people onboard. Replacing the captains' senses and ability to interpret the situation is not an easy task. Again, assuming nice weather and only large ships as objects to be recognized makes things rather easy. But in coastal areas there is a lot of leisure traffic, which might not be seen on radar, especially in bad weather.

A fundamental challenge in autonomous shipping is not only technical, but ethical. At the moment, every seafarer, professional or private, goes to sea with an assumption that they have right to get rescued by other boats and ships in case of emergency. Vice-versa, every seafarer has responsibility to rescue others who have emergency situation. The whole international sea-rescue law is based on this simple principle. When talking about autonomous shipping, one of the most difficult debate will be whether or not we allow that there are ships which do not help the ones who need help.

TYPICAL INHERENT TECHNOLOGIES

Typical inherent technologies and research areas for autonomous shipping are AI/control, computer vision, sensor technology, prescriptive maintenance, automation of rescue operations, connectivity, remote maintenance, and industrial IoT. Except for the Industrial IoT, the technology readiness level for enabling truly autonomous operation of ships for the critical technologies is TRL 2–3, not much higher.

AI/control

In autonomous shipping, AI can play major role in the future. When looking into the navigation, control and decision making, AI can be utilized at least in situation interpretation and determining the optimal decision/control action. In addition, handling of uncertainty and learning with low amounts of data are emphasized as general challenges which relate to utilizing AI in shipping. Let us take a closer look at these.

Situation interpretation: For optimal decision making and control, a crucial part is situation interpretation. This includes fusing information from all sources; radars, lidars, cameras, sound sensors, maps and similar to recognize all visible static and dynamic objects, and to track and predict their movements and motions. One needs to create as good a picture of the situation as possible and be able to evaluate the uncertainty of different sources of information in order to provide good input for control and decision-making systems.

Optimal decision/control action: In the context of autonomous shipping, the technology readiness in navigation, control and decision making in difficult conditions and traffic situations is on a low level. Ship navigation and decision making is defined by the international regulation of avoiding collisions at sea – COLREGS. These "rules of the road" define how to act in situations when two or more ships are on a colliding course.

Handling of uncertainty: This is related to both of above points, but should be emphasized separately. The human decision-making process takes into account the uncertainty of the situation and if, for example, the visibility is low, the human operator will typically be more careful in operation and keep longer distances to other ships as well as decrease speed if possible. In addition, the decision making can be based more on experience if the "sensory information" is not very reliable or accurate. In other words, humans can assess the reliability of the different sources of information and decide to rely more on those that are reliable. Proper handling of uncertainty in a "Bayesian way" is not done in the AI technologies today.

Learning with low amounts of data: When comparing automotive industry with ships, one of the major differences is that the number of "interesting occasions" for the AI is significantly smaller in ships than in cars. The amount in cars is 100–1000 times higher and cars operate mostly in cities where something is happening all the time. Ships are low numbered, speeds are significantly lower and the occasions on how often the ship needs to make decisions are very rare. Therefore, the learning algorithms which interpret the situation and classify the other ships, for example, will need to work on significantly lower amounts of data than on cars. One needs to be able to combine training data from several ships as well as possibly generate data using simulations, at least partially.

Computer vision

Computer vision is connected to the sensor technology and AI but is emphasized separately here as a system solution which will combine the sensory data and AI to create the best possible overview of the ship surroundings.

Prescriptive maintenance

Prescriptive maintenance of the critical components is needed when going towards unmanned and autonomous shipping. Among the classification societies, there are discussions in the industry of requiring that there should be a failure prediction system for all critical components which can predict failures several hundred hours in advance. Prescriptive maintenance can also be a topic where AI can be utilized.

Sensor technology

Sensor technology is developing very rapidly due to significant push from automotive industry. The prices of lidars, low/mid-range radars, infrared sensors, and cameras

are rapidly decreasing. The development of solid-state lidar technology will decrease laser-scanner prices to a fraction of what they are today. That will enable increased coverage and redundancy of sensor systems in a cost-efficient way. The challenge is that the navigation radars are not designed for unmanned operation. The officer of the watch is responsible to tune the radar parameters based on the environment and that affects significantly on how well the radar is able to recognize the environment. When it comes to other technologies such as lidar, microwave radars, cameras, and IR cameras, the harsh conditions at sea will have significant impact on the reliability and durability of the sensors. The sensors are typically designed for automotive industry and therefore the typical ranges of the sensors are not enough for shipping. The development of situation-awareness sensor technologies for shipping is still a challenge.

Underwater situation awareness presents additional obstacles. The navigation maps are not always very reliable and therefore continuous mapping of underwater environment might be needed in order to operate the ship reliably. Technology readiness for underwater sensor technologies for continuous mapping (in a commercial scale when it comes to cost) is not high enough.

Automation of rescue operations

A major obstacle when it comes to unmanned/less-crew shipping is the rescue operations. Currently all ships have responsibility to do rescue operations if there are other ships or people in need of help. This requires typically at least five people: one controlling the ship, one working on the deck with the life boat, one controlling the life boat and two people lifting an unconscious person from sea. Increasing the level of automation in rescue operations is needed before ships can operate longer distances without crew. This can be a research topic on its own.

Connectivity

It is anticipated that 5G technology will come to market by the end of the decade. That will enable significantly higher bandwidth with guaranteed latency of milliseconds. This is a breakthrough for autonomous and remote operation of ships on coastal areas. Of course the 5G technology cannot be used outside the coastal areas, but on the other hand, it is mostly in the coastal areas where very high-activity operation takes place.

In addition, satellite technology is developing fast. There will be more bandwidth with lower prices available. This enables better connectivity at open sea. Also, there are companies such as ICEYE[1] working on building networks of microsatellites which could enable better connectivity even in polar areas. Similarly, there are companies working on building HF-based connectivity to build satellite-independent connectivity and localization, such as KNL Networks[2]. The breakthroughs in connectivity will enable more reliable and redundant remote operation, navigation and localization.

Remote maintenance

When going towards unmanned ships, the maintenance operations without people onboard is subject to research. Remote supervision and maintenance could be done using robots or drones for various applications.

Industry 4.0/IIoT

The digitalization in shipping has advanced significantly during the recent years, but the general system-level integration suffers from lack of standardization. However, I do not see this as a major research question.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

Probabilistic AI (for dynamic systems)

Generally combining uncertainty of inputs and parameters and taking that into account in the model predictions is not feasible today. A general modelling toolkit with efficient implementation is missing. This also includes the possibility of handling dynamic systems in a proper way. The commercially available solutions are not very good at that.

Combining prior information to AI

So far, most of commercially available machine-learning toolkits are blackbox modelling solutions. This means that one cannot bring prior information to the models very easily. Prior information could decrease the degrees of freedom and therefore the amount of parameters needed to be learnt from data. This would enable training the models with less amount of data.

Combining AI and MPC

The existing toolkits for MPC are not able to utilize machine learning, or AI in general, for (nonlinear) optimal control. Combining data-driven modelling and MPC in a way that it can be applied in real problems is not existing today.

All-weather day and night computer vision

When AI is used to classify and recognize objects such as ships, one typically uses mainly camera data for that. The challenge with ships is that they operate day and night at all weather conditions. This means that if the AI is trained using day-camera data, it does not work at night. One can, of course, use IR cameras for this purpose, but they require a separately trained model. This means having separate training data sets for sensors such as IR cameras which are optimized for night conditions. A breakthrough would be to enable using the same model for day and night conditions.

Stochastic, scenario-based multi-objective optimal nonlinear control and decision making

The traffic situation is always multiobjective and stochastic. In navigational decision making, one needs to assess the uncertainty of the information including the knowledge of other ships, the depth of the water, sea currents,

wind, and similar – and be able to do risk analysis such as "what if the other ships are not following COLREGS?".

Reliable (high-bandwidth) long-range global connectivity

Currently, when going beyond the reach of commercial 3G/4G networks, the main data connection to the ships is commercial satellite connection. These connections are very expensive and not always 100% reliable. Therefore, in order to enable autonomous shipping, the connectivity to the ships needs to be improved.

GNSS independent global positioning

As described previously, a global positioning system that is independent of the GNSS satellite system is needed for reliable global positioning.

Self-healing/reconfigurable/fault-tolerant systems

The power train and other systems onboard the ships today are not designed for unattended operation. When the number of people onboard are decreased, the systems need to be more fault-tolerant or even self-healing.

Underwater situation awareness sensors

The current sensors available for mapping the seabed are not commercially feasible due to very high cost, or are not robust enough for commercial shipping. There should be a system based on, for example, sonar or optical measurement which can provide 3D profile of the surroundings of the ship under the water.

POTENTIAL FOR SWEDISH POSITIONING

AI/Control, prescriptive maintenance, computer vision

When looking at the background of the Swedish research and industry as pioneer in system identification and control theory, this is a good basis to build the AI know-how which is crucial for the autonomous navigation and control. This topic has also significant synergies in Swedish automotive, aviation and defence industry where similar solutions are required.

Remote maintenance, automation of rescue operations

Swedish industry has strong background in robotics and therefore the remote maintenance topics could be very relevant. There are strong research groups in this area as well.

POTENTIAL PLAYERS

There are several major industrial players who provide systems and solutions for ships and have publicly communicated about their interest towards autonomous shipping. These include, for example, ABB, Rolls-Royce, Wärtsilä, Kongsberg and GE. In addition, there are smaller companies working mostly on autonomous surface-vessel development.

For the ship-to-shore and ship-to-ship connectivity, the main players are major telecom companies, such as Nokia and Eriksson, as well as satellite companies such as Inmarsat. In addition, there are small companies such as KNL Networks who are developing alternative solutions for connectivity.

NEED FOR SUPPORTING ACTIVITIES

There are several consortia globally for development of autonomous shipping, such as ONE SEA in Finland and NFAS in Norway. The ONE SEA ecosystem in Finland includes main industrial players such as ABB, Rolls-Royce, Wärtsilä and Cargotec. Norwegian Forum for Autonomous Ships (NFAS) involves Norwegian and international companies and universities. There are test areas for autonomous ships at least in Norway and Finland.

One of the major collaboration themes for all big players in the field of autonomous shipping is to facilitate the development of international rules and regulations and classification-society rules to keep up with the development of the technological solutions.

There are also major industrial boards which work together towards the development of a more modern rulebase which would take into account the possibilities offered by new technologies.

One supporting activity that could help in the development of autonomous shipping is the enabling of academic collaboration across the disciplines. The marine engineers today are typically still very much focused on traditional marine engineering, including ship stability, hydrodynamics, and similar. While this continues to be very important, the requirement for advanced control engineering, data-driven modelling, optimization, software development, and similar, are skills that very often are not mastered by the same individuals who master ships. Having more cross-disciplinary research and education will facilitate development of autonomous shipping.

In addition, in order to facilitate the actual development, test and research platforms are needed. This means to have actual ships (or boats) that can carry realistic control systems and sensors onboard and enable testing real solutions in sea conditions. Such test platforms could, at least partially, be publicly supported.

Footnotes and links:
1. www.iceye.com
2. www.kyynel.net

Unmanned aviation

Gunnar Holmberg, Saab, gunnar.holmberg@saabgroup.com

OVERVIEW

The first hundred years of flight have had a huge impact on travel and trade. From technical and operational points of view, giant leaps have been taken in terms of safety, productivity, regularity and similar. Many aviation-related technologies have dealt with safe operations based on a pilot on board, responsible for the safe performance of the flight.

Recent progress in technologies like communication, sensors and electronics have made it less obvious that a pilot needs to be on board. Consequently, unmanned aviation with the pilot operating the system remotely has been central to the aviation innovation in recent years.

When there is no pilot on board constraining the size and behavior of the system, aircraft can suddenly be made much smaller, lighter and simpler than before. Technologies for unmanned aviation will likely change the pace of development in manned aviation and contribute to improved safety, environmental friendliness and efficiency.

Typical application areas for civil unmanned aviation can be seen in medical, agricultural, forestry, freight, security and similar areas. In the long run, unmanned aviation could contribute to, among other things, more efficient and sustainable transport systems, precision agriculture with less environmental impact, novel low volume cargo networks and distributed healthcare for sparsely populated areas improving quality of life in such areas, better safety and security dealing with potential disasters caused by natural disaster, human intervention or other.

INTERNATIONAL MATURITY

Unmanned aviation is regularly applied for security and surveillance purposes globally today. This is a beginning of the application of unmanned aviation with a possible continued development in line with possible scenario in the fact box.

A POSSIBLE SCENARIO FOR UNMANNED AVIATION

2017

- UAS are applied for a majority of flights in ongoing military conflicts.
- UAS are in experimental use for a number of state and civil applications such as security, disaster relief and environmental monitoring.
- The fraction of unmanned aviation in total aviation business is in the order of single percents.

2040

- People have started to hesitate to fly with aircraft needing more than one pilot on board, since it is considered a sign of lacking safety in the system.
- Global cargo networks operate the modern part of the fleet unmanned for competitiveness reasons.
- Regional cargo networks for sparsely populated areas are operated by unmanned aviation.
- Business jets have started to operate without pilot on board but the option to bring the pilot is still offered.
- The estimated fraction of unmanned aviation in total aviation business is in the order of 15–30 percent.

2170

- People are worried to go with aircraft that rely on a pilot on board.
- Pilots are on board mainly in legacy aircraft and when needed for particular reasons in all types of aviation.
- The estimated fraction of unmanned aviation in total aviation business is in the order of 75–90 percent.

Today, Amazon, Google and others are looking at package delivery, with a focus on rapid delivery in urban environments. Concepts exist for personal transportation such as single-person air-taxi quadcopters. Plenty of research efforts are performed for agriculture, environmental studies, forestry, public safety, wildlife protection and many more areas throughout the world.

CHALLENGES

The safety contract is currently being negotiated; regulations are immature, constraining the possible usage. When pilots and passengers are on board the aircraft, safety is derived from the risks that the people on board are exposed to and it is then assumed that third-party risks are orders of magnitude less than for those on board.

Further, it is assumed that the rationale for performing a flight justifies the risks, as those involved often are the same people exposed to risks and rewards (for instance passengers). This does not translate directly to unmanned aviation, where third-party risks are dominating. The approach so far is to apply manned aviation principles and safety levels for larger unmanned air vehicles, while smaller vehicles operating at low altitudes are different and new principles are being applied. The main technologies needed to compensate for the pilot contribution to safety are collision avoidance (known as detect-and-avoid), response to systems degradation and handling the limited reliability of communication links.

Today's regulations are based on a responsible operator (pilot) that "shall always be in command". This does however not necessarily imply that the pilot at any stage needs to have a real-time possibility to affect the system – except the possibility to override. Strategies to hand back the system to the operator in case the system fails to operate successfully is not expected to give a substantial safety contribution, it might even counteract the safety case.

Many applications for unmanned flying systems that include sensors in one way or another have obtained an

initial foothold, for instance aerial photography, geographical mapping and cattle counting. Following this, a debate on personal integrity and surveillance has flourished. Currently, this has led to an application of camera-surveillance laws in Sweden that temporarily has inhibited progress of photography-application areas. Attempts to overcome this demonstrates some challenges; the approach to ensuring personal integrity under the use of on-board sensors is a key aspect to resolve in society to progress further.

Standards and regulations currently developed with active contributions from Sweden. Both Europe and US pay high attention. It is important to note that the approach of these efforts align with current regulations for manned aviation, as defined by ICAO, for unmanned systems flying in airspace used by other aircraft.

As seen in several areas like autonomous driving and Industry 4.0. there are huge expectations globally for the importance of mastering autonomous systems technologies to remain competitive and gain new evolving markets. This has led to extensive investments from industrialists, venture capitalists, research foundations and similar. In the case of unmanned aviation, large investments have taken place globally, and in particular the US have established a technology, operation and regulation foundation based on military acquisitions and operations working closely with the research community. This is followed by many efforts to develop unmanned-aviation solutions contributing to the potential application areas.

When the regulatory obstacles for traffic insertion together with manned aviation and integrity aspects are resolved, there will be challenges for small nations like Sweden, both to find the right niches and to find resources (financial and knowledge) to be competitive and then be able to gain strong business in the global competition.

TYPICAL INHERENT TECHNOLOGIES

Unmanned aviation could be seen as an enabler for innovation in various application areas. The contribution from unmanned aviation is typically flexible positioning of a sensor, or the ability to transport things as well as humans in the future.

Avoiding the sizing constraints of having a pilot on board opens up for many new affordable applications in the future, not least since unmanned aerial systems constitute a convenient platform for sensors of different types. Technologies contributing to these applications are probably more related to the application development than to the unmanned aviation platform itself.

The areas expected to be most important are:

- **sensor development** – an area which sees rapid progress and will continue to enable new applications;
- **communication** – meeting stricter requirements on robustness, resilience, latency and similar which simplifies the use of remotely controlled systems, and in which more bandwidth will continue to expand applicability;
- **precision positioning and timestamp, navigation** – enabling for instance fusion between platforms opening up for increasing scope of applications where several systems collaborate (large events security, disaster relief, precision agriculture and similar);
- **the combination of mixed-initiative and sliding autonomy** for several collaborating systems would increase the reasoning between systems and humans enabling the ability to perform complex tasks with a flexible use of the systems (a typically demanding situation is a rescue operation, but could also play a role in for instance agriculture where farmers are seen to use the systems for diverse applications);
- **expanded approaches** to safety, security and resilience.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

The most important short-term breakthrough would be the ability to design a system with sufficient autonomy to have a defined safety level accepted by the public through regulations. This contains short-term gains such as standardization based on current technology, operation and liability principles establishing a platform where additional technologies could contribute with safety and resilience. Such technologies range from better sensor technologies increasing margins to novel principles of achieving safety using learning approaches.

It might be that in the longer run, such changes would create novel foundations for achieving systems with safety beyond what is achievable with current approaches that actually tend to inhibit the application of novel technologies not meeting classical criteria of deterministic behaviour.

Also enabling would be a breakthrough in smart heterogeneous interacting systems. As the smartness of systems are growing and applications are increasingly interacting, the role of the human in interaction with the system is changing. We would expect novel paradigms for interaction between humans and systems with a varying degree of authority.

Systems containing this level of complexity in the interaction with users including almost infinite number of states and varying authority over the system require novel approaches both to condensing the state to an interpretable level of information over time and being able to suggest various actions and consequences.

Cognitive companions combined with normal visualization approaches offer promising opportunities. Even more intriguing is when big-data analytics and learning starts to create insights that are not foreseen or easily explained to humans interacting with system. An early example could be information gathering in precision agriculture that adapts depending on expected information need. At the same time, it is important to offer full transparency to the operator of the rationale behind the adjusting behavior.

POTENTIAL FOR SWEDISH POSITIONING

Collision avoidance, or detect-and-avoid, is globally recognized as a key technology to enable manned and unmanned aviation to operate together unconstrained. This

technology is currently at TRL 6–7 and Sweden has a long history of developing systems for manned aviation avoiding both ground and air collisions. This has positioned Sweden as leaders for the now successfully completed European MIDCAS project gathering key European actors. A follow-on project is ongoing, with Swedish lead, and the project is globally recognized. This offers an opportunity for Sweden to be early providers of unmanned systems that are possible to operate integrated with other air traffic on equal conditions.

Swedish industry have products with sophisticated user situations, including extremely demanding user environments, like fighter-aircraft cockpits where the Gripen cockpit is recognized for its strong support in complex situations. This combined with research on interaction, mixed initiative, cognitive companions, task negotiation, big-data analytics and more offers opportunity for disruptive progress in interaction with heterogeneous systems of systems.

The growing complexity of systems and the concerted operation of systems together increase the need for efficient development testing and maintenance; not least the introduction of smartness and learning increases challenges in terms of the ability to design systems with predicted level of safety, reliability, resilience and similar. The Swedish system-building industry combined with research strengthening these aspects should make the Sweden suitable for hosting these types of systems. Challenges as discussed above relating to funding and knowledge access remains.

POTENTIAL PLAYERS

As mentioned, Sweden enjoys strong aviation, communication, sensor and systems industries complemented by research groups with potential strength in many of the dedicated areas. From this, we assume that Sweden has a good starting point for providing both unmanned systems and application systems designed for a certain set of applications. The actors currently involved are likely to be able to play a role in the future, including for instance the many start-ups gathered in the Swedish Aerospace Cluster.

Autonomous driving of vehicles could probably benefit from the structured approach to safety that aviation has taken. Swedish automotive industry including suppliers has opportunities here. Many industries increasing automation for vehicles or industrial systems with demanding requirements on safety and reliability have similar opportunities. If this proves to be true, then we could see possible contributions of various importance to a large portion of Swedish system-building industries.

NEED FOR SUPPORTING ACTIVITIES

Most of the potential areas described above make sense at a high level of system integration and complexity. This often causes challenges in finding relevant research approaches as well as bridging over to industrial relevance. From experience, we have seen that demonstrators facing progressively more challenging demonstration goals offer help to focus research, to connect to relevance in applications and to continuously mature technologies/approaches towards application.

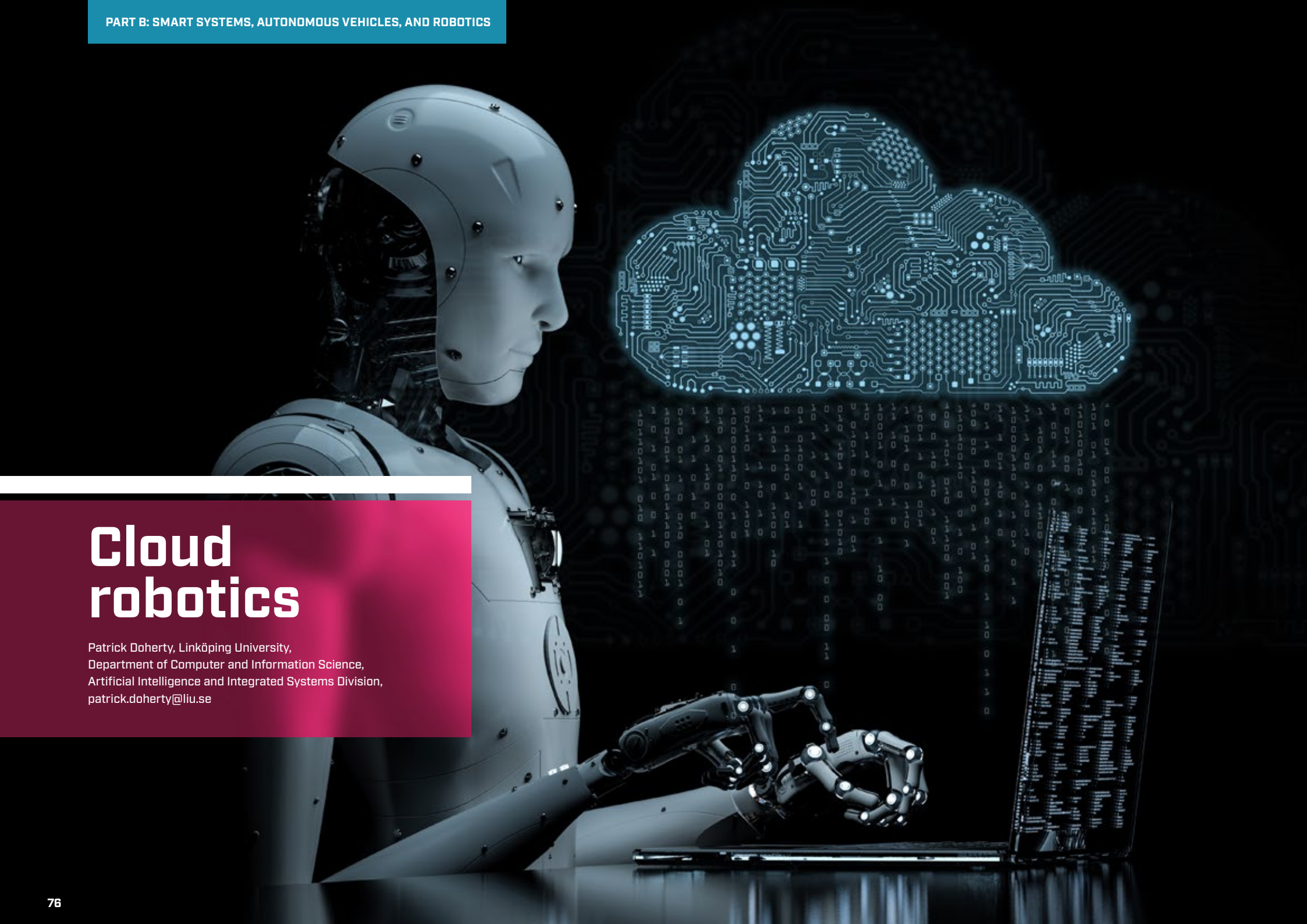
This is the rationale behind the demonstrators suggested in WARA. These have the potential to progress many application areas within unmanned aviation and other areas. For example, WARA-PS is designed to use a multitude of heterogeneous vehicles to solve a rescue mission that is initially vaguely understood and should be continuously re-planned as the situation is becoming clearer and resources are added. WARA-PS was built for the rescue scenario, but other applications include other disaster scenarios like forest fires, earthquakes, tsunamis and so on, and also slower applications like critical-infrastructure monitoring, large-event monitoring, precision agriculture and forestry. These all face problems that could benefit from solutions being the focus of WARA-PS.

When initial scenarios have been demonstrated in WARA-PS, several routes could be entered to increase demonstration challenges reflecting the research directions identified and selected for WASP in the longer term.

As mentioned above, public acceptance is sometimes the most challenging part of when progressing towards application of new technologies. The key issues currently discussed in relation to unmanned aviation is dominated by three aspects: safety, integrity and security.

- **Safety** is probably best supported by a structured approach and gradual introduction of applications including demonstrations and early small-scale adoptions with constraints on where to fly.
- **Integrity** is an interesting debate in which unmanned aviation is just a small part of a situation where consensus has to be achieved, regarding how to use technologies to reflect a general perception of societal benefit that is superior to the perceived integrity risks. The current drone-camera debate in Sweden illustrates this. It might be that in addition to supporting the debate to arrive at a conclusion reflecting a society-preferred perspective, it is important to highlight technology's possibility to actually support integrity by, for instance, erasing faces in images when the portraits are not explicitly stated to be the purpose of the picture/film. The European Commission has developed interesting scenarios that perhaps should be taken to Sweden to increase the precision of the discussion here as well.
- **Security** is a continuous challenge, where it is important to ensure both that a system is not attacked by malicious actors and that a system cannot be used for malicious purposes.

Finally, due to the high international attention to these topics, it is probably important to have a strongly international approach when looking at venture capital approaches, technology and market alliances.



Cloud robotics

Patrick Doherty, Linköping University,
Department of Computer and Information Science,
Artificial Intelligence and Integrated Systems Division,
patrick.doherty@liu.se

DEFINITIONS

Cloud robotics: A field of robotics that attempts to invoke cloud technologies such as cloud computing, cloud storage, and other Internet technologies centred on the benefits of converged infrastructure and shared services for robotics.[1]

Cloud Robot and automation systems: Any robot or automation system that relies on either data or code from a network to support its operation, meaning that not all sensing, computation, and memory is integrated into a single standalone system.[2]

ROS: Robot Operating System, an open-source set of software libraries and tools that can be used to develop robot applications. It can be viewed as robotics middleware in that it provides operating-system-like support for robotics architectures.[3]

OVERVIEW

The term cloud robotics was coined by James Kuffner in 2010, although many of the ideas associated with the concept existed in one form or another as far back as 30 years ago in areas such as networked robotics and tele-robotics. The definitions in the fact box provide a feel for the ideas included in this umbrella term. Cloud robotics is a new way of thinking about robotics. It is an emerging paradigm shift.

The deceptively simple and broad definition of the area has wide repercussions for the future with potentially high technological, economical and societal impact. It is multi-disciplinary in nature and includes the integration of other "contexts" such as deep learning, cooperative learning, big data, semantic web, automated planning and Internet of things. Additionally, it has the potential to leverage crowd-sourcing but with an additional twist: robot-sourcing!

Here are some examples of current research that exemplify the type or research associated with this context:

Google Smart Cars. These self-driving vehicles leverage huge amounts of data from the Cloud. They index maps and images collected and updated by satellite and Street-view as well as information gathered through crowd sourcing.

Kiva Systems. This company was recently purchased by Amazon in 2012. It is has been renamed Amazon Robotics[4]. It is in the business of warehouse logistics and has automated large parts of the logistics process through the use of pallet robots. The robots themselves are relatively lightweight in terms of computational and sensing capacity, but they communicate wirelessly with a local server, coordinating, planning and updating local changes in the factory floor environments.

Cloud-based robot grasping. Detection, identification and grasping of known and unknown objects in cluttered dynamic environments is one of the more challenging and computation intensive tasks in robotics. Robust solutions to this general problem would have tremendous impact in many application areas where robots are intended to be used as service robots, in emergency rescue and in agriculture, among others. Kehoe et al[5] describe a cloud-based robotic system that leverages open-source software and Google's Object Recognition Engine where the cloud-based object-recognition system is trained using big data in the form of multiple images to generate an understanding of many different objects. A robot can then associate its sensor data with this repository in determining what kind of object it is looking at and determine what might be the best way to grasp it.

RoboEarth. RoboEarth[6,7] was a highly publicized and successful EU project first announced in 2009. It was completed in 2014 and the ideas have continued to be developed in follow-up projects such as RoboHow[2]. The vision was a world-wide web for robots. The idea was to develop a networked and distributed database repository that could be leveraged and shared by robots to enhance learning, planning and reasoning processes.

These examples are representative of some of the ongoing activities in this area and such activities are only growing. Note that examples from automation are not covered here, but the combination of cloud and industrial automation are having just as much if not more current technological impact as described in Industry 4.0 and other such endeavors.

Cloud robotics represents a confluence of several different research areas and technologies and the benefits are numerous. Here are a few:

Access to data and resources

The cloud has the potential to provide robotic systems with access to huge amounts of data and resources for computing with that data. Many of the processes necessary for robust, resilient and safe use of robotic systems are highly computation- and data-intensive. They involve large collections of images, videos and sensor data that needs to be fused and processed for different purposes such as localization, mapping, object detection, semantic labelling and similar.

For many of these processes, doing this efficiently on-line is currently beyond the processing capacity of stand-alone systems. Additionally, current success with robotics has a great deal to do with the use of big data and machine-learning techniques where the data does not fit into conventional database systems. Imagine if one could unleash both the massively parallel processing capabilities, distributed data collection, massive storage capabilities and the smarts and intelligence of many humans and other robotic systems in a principled manner for use by individual robotic systems? Techniques such as cloud-based bundle adjustment, sample-based Monte-Carlo analysis, comput-

ing robust grasps in the context of object uncertainty and SLAM, among others, could be made much more efficient and robust. In fact, through learning processes these tasks would only get better.

Collective learning

The cloud offers an ideal infrastructure for collective robot learning. Large collections of robots could share experiences in the form of trajectories, initial and desired conditions, control policies and data on resulting performance and effects of task execution. This would provide a basis for shared path planning and improving the capabilities of robots limited by computational capabilities, such as micro-drones. Robot learning would be greatly accelerated by this approach. As Ken Goldberg states in an interview: "Putting it simply, one robot can spend 10,000 hours learning something, or 10,000 robots can spend one hour learning the same thing." Of course, one would have to figure out how to fuse and share the results.

App repositories

There is a growing repository of robotic know-how in the form of ROS packages and other algorithms developed by diverse research groups. ROS is the de-facto standard for sharing software among robotic research groups around the world. Imagine developing an international system analogous to smartphone-app repositories, where apps for robots are provided on-the-fly when needed? The potential is fascinating and the impact would be huge in terms of accelerating the capacity of many robotic systems and shortening the research time required to develop robust robotic systems.

INTERNATIONAL MATURITY

As stated previously, research trends leading to the modern version of cloud robotics have existed for over 30 years, beginning with the use of networks with robotics, tele-robotics and the evolution into leveraging the World Wide Web for Robots, where RoboEarth is a prime example.

But in the past ten years, there has been an acceleration of these ideas into the modern version of cloud robotics due to the advent of Google, the maturation of the Cloud and the qualitative breakthrough with applied machine learning and deep learning due to increased computational capacity and access to large amounts of data.

The confluence of AI with robotics has also led to a demand for such an infrastructure and practical leveraging of AI technologies in robotics systems is only in its infancy. "AI technology" is intended to be used in the broader sense here to include topics other than machine learning such as automated planning and knowledge representation, speech recognition, agent-based paradigms and cognitive computing. AI is not just machine learning as many journalists like to portray it.

Leading players are the major technology companies such as Google, Microsoft, IBM, Amazon, Apple and Facebook which are essentially morphing into AI companies. Google alone has purchased about a dozen robotic companies in the past few years and has been a major player in Willow Garage[a] and the development of ROS. Additionally, a large number of top robotic groups at leading universities have activities in this area and are combining this with interesting forms of research with Q/A systems and human-robotic interaction. The latter combination of topics is yet another important context that would require its own foresight.

Although there is some maturity in the area and the infrastructures required for research are somewhat in place, the topic and technologies surrounding it will be under intensive development in the coming decades. The belief is that this approach to robotics represents a paradigm shift of immense proportions. Knowledge is power and the cloud potentially offers this to robotic systems if it can be standardized, harnessed and applied appropriately.

One can imagine that the first robust commercial uses of many ideas associated with cloud robotics will emerge with smart car technologies. Google has already shown the potential and benefits of doing this. Another interesting area ripe for application is with smaller drones that lack adequate on-board computational capabilities and memory storage.

CHALLENGES

There are many challenges associated with this context. Here are a few:

Latencies

Current cloud infrastructures have problems with varying network latencies and quality of service. Time-sensitive robotic processes would require improvement in these areas. Solutions to deal with latencies would also influence how one develops robotic architectures themselves. For example, any-time approaches to algorithmic processes would be important. An additional idea here would be to move to more robust communication protocols such as 4G and 5G. New algorithms dealing with latencies and load-balancing will have to be developed to ensure improvement in this area.

Protocols

Cross-platform protocols for representing data need to be developed if large numbers of robots are going to share different types of data. There is already some activity in this area such as how to represent geographic date or point clouds, but there is much left to do here. For example, what would be the proper data format for trajectories or streamed data in order to maximize sharing of data?

Noise, duplication, consistency

Combining many different data sets from distributed sources is a huge problem in terms of varying degrees of noise, duplication, consistency and similar. This is in fact still an open problem that arises in some sense with the semantic web and distributed databases where some solutions could be borrowed and expanded upon, but there are additional facets to this problem that arise in this context.

For example, much of the data being discussed is sensor data; dealing with varying noise in the data becomes important.

Sparsification

Sparsification of data into efficient representations at various levels and processing at various levels will be a significant challenge. For instance, one will not want to pass around point-cloud data given limited-bandwidth problems. New algorithms are also needed that scale to the size of big data.

Scalable inference and reasoning techniques

Inference and reasoning techniques developed in the area of AI/knowledge representation have great potential here, but these approaches would have to be scaled to highly distributed systems with huge amounts of knowledge. The data and information in the cloud would have to be put in the right formats so reasoning could be done distributively and with heterogeneous data types. For example, suppose a service robot is asked to serve coffee. It can find a coffee machine in its location but does not necessarily know what type of machine it is and how it operates. It should be able to access any type of reference from the cloud in computer readable form in order to reason about that particular coffee machine type and its operation. Scalable stream-reasoning techniques will be of great importance here since much of the data used is temporally tagged.

Security and privacy

Security will be a major challenge. Just as this problem arises in the Internet of things, robots suddenly become some of those "things" and one does not want one's robot hacked, especially if it is a smart car. It has already been demonstrated that this is doable.

Privacy issues are another problem. If robots are passing around high-resolution maps of areas they operate in, there must be a way to distinguish sensitive maps from publically accessible maps.

Planning

There will also be a need to rethink what planning is and how to automate planning in a new way. As an example, if a robot has a motion-planning problem where it is also using actuators to achieve tasks and it has access to 10,000 similar trajectories in similar contexts generated by other robotic systems, how might that data be leveraged in the generation of a new motion and task plan in a timely manner?

Additionally, one will want to push the idea of knowledge-intensive planning where a world model used by planners is not just implicit in the action-state specifications, but may involve use of more abstract world information and involve inference on-the-fly. New planning techniques would be required that deal with robot-to-robot cooperation and human-robot interaction, both in the plan-generation and plan-execution phases.

As one might understand from some of these challenges, many of the traditional topics in robotics and AI would have to be rethought in this context. Additionally, integration of the disciplines of AI, robotics, control and computer science become paramount.

TYPICAL INHERENT TECHNOLOGIES

We have already referred to many of the technologies/research areas involved in cloud robotics. Certainly, one of the most important is the integration of AI with robotics. This has been happening gradually, but not fast enough. In the '60s and '70s, AI and robotics were the same areas. Or at least there was great overlap. This changed in the '70s and '80s as AI started to focus on higher-level cognitive capability while robotics focused more on bottom-up approaches to localization, sensing, navigation and gripping.

It is only more recently that we are beginning to see attempts at deeper integration of technologies from both areas using both probabilistic and logical techniques. This context is not an accident and is due in large part to maturation of the technologies, access to greater computational capabilities and access to large data sets for learning.

Additionally, solutions to many problems in AI/robotics require both higher-cognitive and lower-cognitive capabilities, just as is the case in humans. Base robotic platforms are now at a point where these integrative technologies can be tackled and they provide added value. In fact, it is difficult to conceive that robust, resilient, safe robotic technologies operating in uncertain environments will happen without these integrating confluences. The interesting twist with cloud robotics is that one is advocating the distributed construction and leveraging of a "collective brain" in the cloud that can be accessed by individual robots anywhere that will accelerate their capabilities. By "collective brain" we mean the sum total of knowledge and services generated by many systems and made accessible to any individual robotic system on-the-fly.

The use of the cloud in this context naturally introduces many of the conventional issues researched in computer science. In general terms, the goal here is to get the right information to the right place at the right time and in the right form, so that cyber-physical systems can process information to make appropriate decisions for themselves and/or provide decision support to humans. Research topics associated with this are more at the forefront of current research since they involve almost any IT system being built today. Research revolves around scalable, networked, distributed information systems and efficient communication and bandwidth between nodes.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

Deep learning and Bayesian learning are only as robust as the data sets used as input. One can generate impressive results using these techniques, yet they are still somewhat brittle in the sense that they are based on correlation and not causal models. There is an awareness of this problem and initial research is being pursued in combining semantic models with these techniques.

Doing this would result in more robust and resilient robot behaviors. One can anticipate breakthroughs in this area due to the cloud infrastructures that provide big data at all levels of information abstraction and implicit causal relations between such data waiting to be leveraged, combined and mined.

Due to intensive research with next-generation communication infrastructure for smartphones such as 4G/5G, one can imagine that latency and quality-of-service problems associated with cloud robotics will qualitatively improve, especially due to pressure and experience with the vast group of customers using smartphones.

The semantic web has the potential to offer huge amounts of computer readable, crowd-sourced semantic information about just about any concept one can imagine. Leveraging this semantic information in the cloud with collective machine learning will rapidly improve the quality of learned information in addition to accelerating use of such information for inference in robotic systems. This breakthrough would come by combining research in AI/knowledge representation with machine learning and the cloud.

POTENTIAL FOR SWEDISH POSITIONING

Sweden has some ideal opportunities to create positions in this area:

- First and foremost would be both industrial and academic activity revolving around smart cars. Currently, much of the focus in research seems to be on the smart cars themselves, yet there are tremendous opportunities to leverage cloud technologies in combination with smart cars to provide better understanding and situation awareness of the environments smart cars will operate in, their localization, in addition to dynamic information of traffic, weather, congestion and similar.
- Ericssons is developing cloud technology together with 4G/5G networks. There is an obvious possibility to create a position here. In fact, Ericsson has started late in this area and has a technology in search of applications. There is a reasonable amount of robotics activity here in Sweden, so the combination of technologies offers an appropriate way to progress into this area.
- Cloud robotics is not only multi-disciplinary but involves a great deal of integration activities from diverse research areas. Systems of cyber-physical systems and the inherent software required for integration offer great potential to create positions here. This could involve both Saab, Ericsson and ABB.

POTENTIAL PLAYERS

The WASP consortium of companies and academic institutions have many of the raw prerequisites to advance this area and it would be strategically wise to do this.

From the research perspective, one of the stumbling blocks is the lack of larger integrative projects within WASP that force multi-disciplinary activities, that have a common goal and that provide not only research-funding support but integrative funding support. This might be solved by pushing the proposed national demonstrators more deeply into the research methodology loop and building larger integrative projects on top of the technology required to be constructed for the demonstrators.

There are many pockets of competence within WASP but what is lacking is an incentive to truly mesh these competences together. For cloud robotics, natural players would include Lund University, Linköping University, KTH, in addition to Saab, Ericsson and ABB. The two WARA demonstrators offer ideal test beds to pursue research in cloud robotics, provided one steers technology support for this into the demonstrators and makes sure that one or more integrative projects are targeted and financed using a demonstrator-based research methodology. Financing should not just be for new PhD students but should allow for mid-level researchers and software and engineering support.

NEED FOR SUPPORTING ACTIVITIES

Cloud robotics is very much about integrated systems research and field robotics. In order to be internationally competitive, one requires a substantial investment in not only infrastructure, but investments in human resources on the software and engineering support side.

The WASP consortium recently visited some of the main players in field robotics, such as ETH, Stanford, Berkeley and Singapore. What is striking about many of these internationally acclaimed research labs is the fact that they employ many full-time engineers offering software and hardware support to researchers to assist in system building and experimentation.

Without this kind of set-up, it is going to be very difficult for Sweden to attain a competitive edge academically in this area. Additionally, one requires more of a moonshot mentality in the multi-disciplinary and integrative research required for this area. For instance, Sweden could be working along the lines of entering into the DARPA competitions in rescue robotics or with smart cars, or the numerous competitions with drones in a more principled manner.

Footnotes and links:

1. en.wikipedia.org/wiki/Cloud_robotics
2. A Survey of Research on Cloud Robotics and Automation. Ben Kehoe, Sachin Patil, Pieter Abbeel, Ken Goldberg. IEEE Transactions on Automation Science and Engineering (T-ASE): Special Issue on Cloud Robotics and Automation. Vol. 12, no. 2. Apr. 2015.
3. www.ros.org
4. www.amazonrobotics.com
5. Cloud-Based Robot Grasping with the Google Object Recognition Engine. Ben Kehoe, Akihiro Matsukawa, Sal Candido, James Kuffner, and Ken Goldberg. IEEE International Conference on Robotics and Automation. Karlsruhe, Germany. May 2013.
6. roboearth.ethz.ch
7. RoboEarth - A World Wide Web for Robots Markus Waibel, Raffaello D'Andrea et al. 2011.
8. www.willowgarage.com

Robotics – in factories and at home

Danica Kragic Jensfelt, KTH, dani@kth.se

DEFINITIONS

Industry 4.0: The current trend of automation and data exchange in manufacturing technologies. It includes cyber-physical systems, the Internet of things and cloud computing. Industry 4.0 creates what has been called a "smart factory".

OVERVIEW

A key aim in both small and large-scale industries today is to achieve high-quality, cost-effective, safe and flexible manufacturing. Robotic technology has transformed manufacturing industry ever since the first industrial robot was put in use in the beginning of the 1960s.

The challenge of developing flexible solutions where production lines can be quickly re-planned, adapted and structured for new or slightly changed product is still an important open problem. Contemporary industrial robots are largely preprogrammed for the tasks, not able of detecting errors in own performance or robustly interact with the complex environment and a human worker.

Existing examples of indoor applications of autonomous systems are fetch-and-carry systems in hospitals, office environments, homes for elderly and museums. It has to be noted that these systems are still mainly of the carry type, meaning that they are able to navigate through the environment but require human interaction for the fetch and object-interaction parts.

Existing outdoor applications include power-plant inspections, disaster-site exploration, safe transportation of people and goods and intelligent lawn mowers. These areas are also promoted by the EU and are a part of the European robot initiative for strengthening the competitiveness of SMEs in manufacturing[1].

In the envisioned future factory setups, independent of factory or home/service applications, humans and robots will share the same workspace and perform interaction with the environment such as object manipulation tasks jointly. Classical robot task programming requires an experienced programmer and a lot of tedious work. Programming robots through human demonstration ("learning by demonstration") has been promoted as a flexible framework that reduces the complexity of programming robot tasks and allows end-users to control robots in a natural and easy way without the need for explicit programming. Thus, one of the main enabling technologies necessary to realize this is the design of a framework that enables the robot to cooperate smoothly with the human, working towards the same goal, which may not be explicitly communicated to the robot before the task is initiated.

Robotic appliances offer significant opportunities for the future both in the private and the professional sector. Although still in their developmental phase, robotic technologies of the future will reach sophistication and the necessary robustness and flexibility. The future robot co-workers and assistants, will be empowering the human workers, increasing the productivity and their skill level. These assistive frameworks will be the core of the future intelligent factory, with open manufacturing workcells where robots and humans share their workspace and collaborate. Some of the systems will operate in a team, implementing the paradigm of "winning by number and networking", advancing the areas of networked and embedded systems.

Industry 4.0 is considered to be an ongoing revolution in this respect and it is enabled by the combination of 3D-printing/additive manufacturing technologies, digitalization and autonomous systems technologies. One of the visions for the future is that manufacturing will be transformed to a service driven even more by the digital and robotics technologies where the market will not be dominated by large industrial manufacturers but by low-cost, on-site manufacturing, also taking sustainability into account.

One of the aspects frequently being discussed over many areas (economy, law, politics) is that the digital and autonomous technologies including artificial intelligence are a threat to the job market. The study of Oxford scholars Carl Benedikt Frey and Michael A. Osborne was one of the first to survey the susceptibility of jobs to the above[2]. This was followed by another study[3] stating that on average 54% of European jobs could be automated.

INTERNATIONAL MATURITY AND CHALLENGES

There has been a lot of investments made in building robot platforms for factories of the future. Here are a few examples from ongoing developments in industry and a list of several important research labs.

Rethink Robotics

The Baxter robot from Rethink Robotics[4] is stated to be a "smart, collaborative robot pioneer". The company motives their development by the future need of low-volume, high-mix production jobs. Baxter is seen as a platform that can be easily reprogrammed for new types of tasks. Some of the important features addressed in the design were to make a robot that is safe by design, that is easily integrated in the existing business, that is trained and not programmed, equipped with many sensors (visual, force-torque).

ABB

Another attempt of designing robots that are inherently safe is also a recent example of ABB's concept robot YuMi. However, internal pilot testing at ABB has shown that an assembly automation with YuMi robots but using a traditional programming approach still takes several months to accomplish. There may be several reasons to this: need for manual programming and code writing, non-skilled users, difficulty to completely understand the capabilities of the robot, and similar. Such long integration times make the payback calculation for many potential customers challenging and also prohibits the dynamic production

setting with rapid changeover of production that many potential customers are seeking for. Thus, for ABB, the next challenge is to develop technologies for a complete integration of a new assembly in as short period as possible (for example, one day).

Moley Robotics

In terms of service/home robots, Moley Robotics[5] promises the world's first robot kitchen by the end of 2017. They foresee that their two-arm design will be easily integrated in a regular kitchen and it is planned to be launched in 2018. One important challenge here will be to enable interaction of and in-hand manipulation of objects and also the ability to do this in various media such as air, water and oil. Making also hardware that can withstand high temperatures may be an issue.

Hanson Robotics

Hanson Robotics[6] has for years focused on building robots that resemble human appearance. They focus on expresivness, aesthetics and interactivity to develop robots that teach, serve, entertain, delight, and provide comforting companionship. Building systems that have two arms, that interact with humans and also interact physically with the world around them is however still a challenge even when huge investments are made.

Willow Garage

Another example is Willow Garage[7], This was a combination of a robotics research lab and technology incubator that was focused on developing hardware and open-source software for personal robotics applications. One of the most important outcomes from Willow Garage was the open source software suite ROS (Robot Operating System). ROS has been adopted by both research labs and some companies and many are still contributing to its development although Willow Garage shut down in early 2014.

To make all these systems more flexible, safer, easy to use and adopt, we need to go beyond designing the classical, highly automated industrial settings where robots are used for precise and high volume production, and develop new manufacturing concepts such as assistive robot co-workers and networks of robots with interaction capabilities.

The envisioned robot co-workers and assistants need to empower the human workers, increasing the productivity and their skill level through safe interaction and collaboration. Such assistive frameworks will be at the core of the future intelligent factory, with open manufacturing floors and work cells in which robots and humans share the workspace.

Such systems need to be context-aware and have the capabilities to interact with the environment and manipulate objects. Learning from human example and self-exploration allows handling the complexity of a typical domestic environment with no predefined models of objects and places.

The same is also valid for the service robots to be used in homes but the challenges are even more severe given that our homes are highly irregular and there is a higher likelihood if interacting with unexperienced users posing also an important safety problem.

PROJECT EXAMPLES THOUGH THE YEARS

- **SARAFun** – H2020, Advanced interfaces and robots: Robotics and smart spaces, From 2015-03-01 to 2018-02-28, RIA funding scheme[8]
- **ROSETTA** – FP7, ICT Cognitive Systems, interaction, robotics, From 2009-03-01 to 2013-02-28, Collaborative project[9]
- **VERSATILE** – H2020, ICT, Innovative robotic applications for highly reconfigurable production lines , From 2017-01-01 to 2019-12-31, IA funding scheme[10]
- **RobDream** – H2020, Advanced interfaces and robots: Robotics and smart spaces, From 2015-02-01 to 2018-01-31, RIA funded action[11]
- **SMEROBOTICS** – FP7-ICT-2011.2.1 - Cognitive Systems and Robotics, From 2012-01-01 to 2016-06-30, Collaborative Project[12]
- **CLOPEMA** – FP7-ICT-2011.2.1 - Cognitive Systems and Robotics, From 2012-02-01 to 2015-01-31, Collaborative Project[13]
- **RECONCELL** – FOF-09-2015, ICT Innovation for Manufacturing SMEs (I4MS), From 2015-11-01 to 2018-10-31, IA funded action[14]
- **WEARHAP** – ICT-2011.2.1 – Cognitive Systems and Robotics, From 2013-03-01 to 2017-08-31, Collaborative project[15]
- **VERITAS** – FP7-ICT-2009.7.2- Accessible and Assistive ICT, From 2010-01-01 to 2013-12-31, Collaborative project[16]

RESEARCH LABS

- Berkeley robotics[17]
- Stanford robotics[18]
- MIT robotics[19]
- EPFL Aude Billard
- German Aerospace Center DLR, robotics[20]
- Max Planck Institute for Intelligent systems[21]
- KTH Centre for Autonomous Systems

TYPICAL INHERENT TECHNOLOGIES

For the overcoming of these challenges, mainly three types of research areas can be identified:
- object interaction and grasping;
- learning by demonstration;
- human-robot interaction and collaboration.

All of these are closely related to each other given that they build on rather similar basic technologies (signal processing, perception, statistics, machine learning, planning). One can say that almost any robotics lab in the world dealing with real physical systems needs to have the ability to address these.

Replicating the effectiveness and flexibility of human hands in object manipulation tasks is an important open challenge in the envisioned contexts. This requires a fundamental rethinking of how to exploit the multisensory data and the available mechanical dexterity of robot systems. In comparison to humans or primates, the dexterity of today's robotic grippers and hands is extremely limited. Contemporary robotic hands can grasp only a few objects in restricted poses with limited grasping postures and positions. The main idea followed by many labs is to exploit sensor fusion: monocular and 3D vision data, tactile and force-torque sensing and to investigate which sensors are most relevant in different tasks.

The programming of an assembly task by human demonstration has significant advantages over offline on manual on-line approaches, with respect to the flexibility and adaptability required for a highly efficient production. Even though the demonstration can be conducted with multiple ways (haptic device, motion capturing), the most intuitive and direct one consists of kinaesthetic teaching of the robot, meaning that the human guides it by direct manual manipulation. The interest of the industry and of the research community for human-robot interaction is continuously rising as the collaborative robots can bridge the gap between hard automation and manual operation. However, the kinaesthetic teaching of the robot is a challenging task with safety and ergonomic issues, since it involves the exchange of forces and energy between the human and the robot.

Another important aspect for learning and collaboration is the need for communication. Performance in social and collaborative situations depends fundamentally on the ability to detect and react to the multimodal social signals that underpin human communication. These processes are complex, with coordination being achieved by means of both verbal and non-verbal information. Several virtual systems have made use of a combination of cues from speech, posture shifts and head movements in order to generate active listener behaviors.

A relevant area of research is face-to-face interaction, where considerable efforts have been spent on finding a way of creating a physical embodiment of a spoken dialogue system that is believable and that is perceived as truly co-present. Recently, a group at KTH has achieved increased co-presence by back-projecting their 3D animated character into a neutral mask, creating a 3-dimensional talking head that occupies physical space in the room and is clearly perceived as co-present by onlookers.

This is one of the systems to be used in an ongoing project named "Factories of the future", funded by the Swedish

MANIPULATION – TWO APPROACHES

Performing planning for manipulation requires a suitable representation. A traditional approach in robotics is to represent the world state in terms of a latent variable. In this way, dynamical systems can be formalized as Partially Observable Markov Decision Processes (POMDPs), which allows reasoning about uncertainty in action effects and state observability. Some works employs the POMDP framework to track pose belief of a known-shaped object and localizes the object by planning grasping and information gathering trajectories. Learning POMDPs from exploration data without suitable priors have only had limited success, because of the curse of dimensionality (of the observations) and the curse of history (of required training sequences). It involves high-dimensional (observation) spaces and for increasingly longer horizon an exponential amount of dfferent sample sequences.

Learning a control policy for a robotic manipulation skill can also be formulated as a reinforcement learning problem. In a basic setup the control policy is a mapping from state of the robot, typically represented by its joint encoder values and also it end-effector position in the Cartesian space, to the joint motor torque commands. The policy is also called the controller. In a reinforcement learning problem the dynamic model of the process is unknown. This leads to two broad approaches to learning policies: model-based and model-free. Industrial robotic arm controllers require extensive human involvement for determining tolerance specification or robotic kinematic parameters and fine-tuning. Reinforcement learning can potentially enable autonomous robots to learn a large set of behavioral skills with minimal human intervention. Practically, however, robotic applications of reinforcement learning often compromise the autonomy of the learning process in favor of achieving training times that are practical for real physical systems.

Foundation for Strategic Research. This makes it possible to develop a robot head than can achieve mutual gaze and shared attention at object in a shared space in which the situated interaction takes place. The group has also developed methods for computer-meditated interaction that makes it possible to manipulate ongoing interactions in systematic ways, in real time, while maintaining a high degree of ecological validity. Although a relatively new idea, incremental speech processing has attracted significant interest in the research community.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

The below is general for all the tree research areas above:

Imagine a setup where a human demonstrates an assembly task or how to prepare a meal and the robot can learn this just through observation, first using cameras to make a general plan but then master the execution by training the task given its own sensory and motion capabilities. Also, the robot may realize it is not able to do something and it asks humans for help – but after performing an extensive search on the web for a solution.

Scenarios like these may be difficult even for humans given that it is sometimes not clear what action one is allowed to take and what not. A number of research breakthrouhgs, in all three research areas, might be needed for this scenario to become reality:

- the ability to extract meaningful information from video sequences in any setting (no special lighting needed or human teachers wearing special clothes, and similar);
- the ability to represent the extracted knowledge such that it can be used for planning/execution by a robot but is also meaningful to humans (humans can understand it) to enable collaboration when needed;
- safe physical human-robot collaboration – a human and a robot performing a task jointly where both dialog and gestures can be used to resolve unclear situations;
- machine-learning methods that learn from "small" data (few examples) but have the ability to learn incrementally when new, meaningful data is available; also, the ability to deal with uncertainty and provide confidence values;
- grasping and in-hand manipulation of any type of objects (different materials, sizes, different media).

POTENTIAL FOR SWEDISH POSITIONING

In order to assess the scientific solutions, a close collaboration with both large manufacturers and SMEs is of extreme importance. One needs to establish new ways of collaboration that ensure a more long-term knowledge exchange such as, for example, internships, workshops and sabbaticals.

In addition, building strong educational ground in WASP areas is of long-term importance. Sweden has historically been strong in building classical industrial companies and one important question with high automation and digitalization may be an increased unemployment rate. This will impact both experienced workers and those that are new at the job market. Sweden can therefore become a leading country that implements lifelong learning initiatives through close collaboration between academia and industry. The areas where this is going to be most important are key areas in WASP: software technology, robotics, artificial intelligence, machine learning.

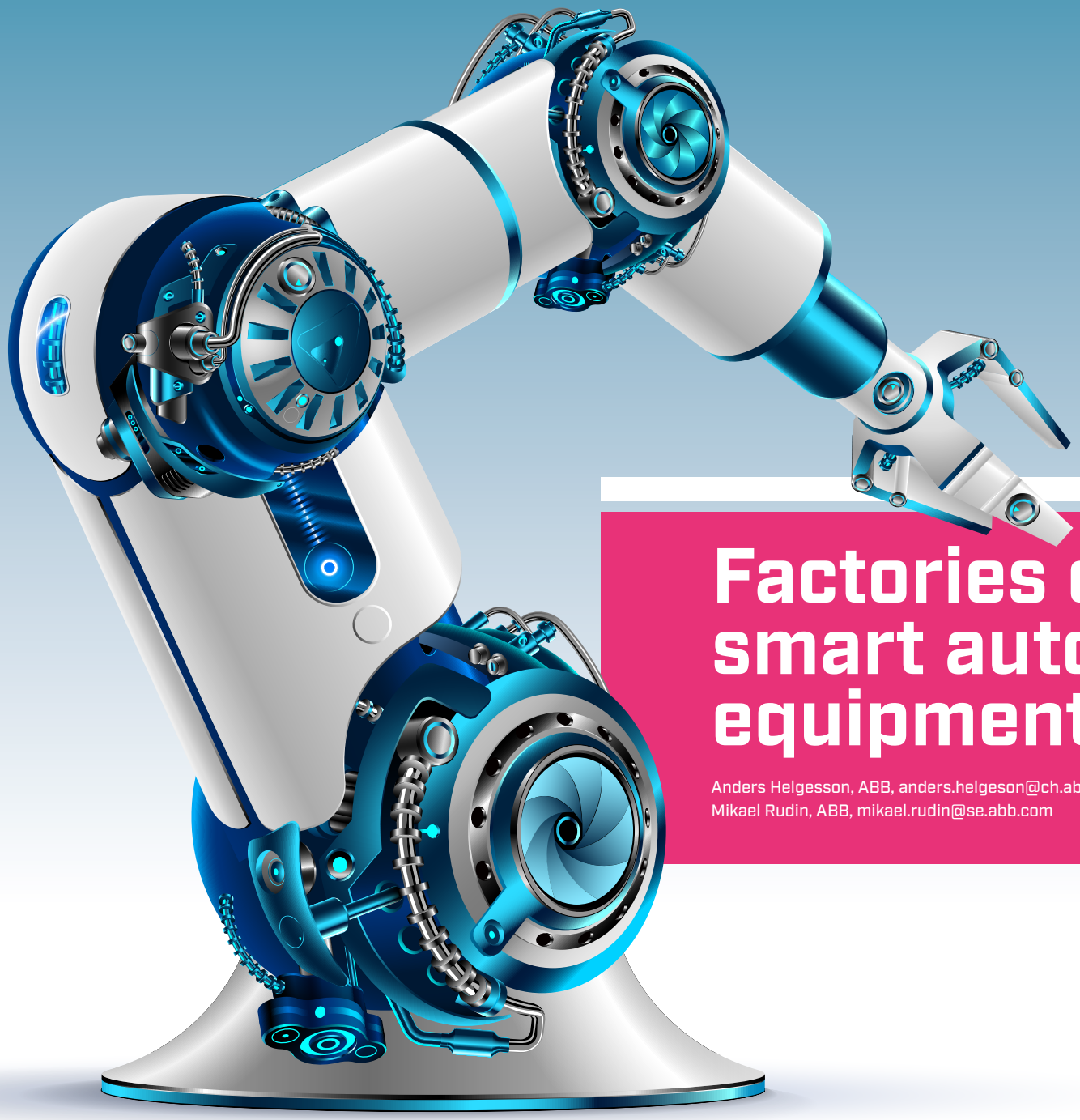
POTENTIAL PLAYERS

WASP is an excellent example but now one would need also "end-users" – companies that are early adopters of ABB's technologies for assembly applications. ABB is one of the companies that are interested in close collaboration with the academia but the collaboration has so far been mostly supported by EU projects, Vinnova and recently by WASP. Electrolux may be antoher interesting player given their early development of the autonomous vacuum cleaner. Husquarna is yet another interesting company from the perspective of autonomous home applliances.

The area of robotics would benefit from a closer collaboration between academia and industry. One of the problems that usually takes a long time to settle is NDAs due to the "rigidity" from both sides. One could build three testbeds (one in a company and two at the universities) with the goal of exchanging the knowledge so that all three sites contribute with different technologies; however, everything needs to be integrated and tested at all three sites by doing, for example, weekly integrations. This would enable an active collaboration between students in the academia and employees at the company.

Footnotes and links:

1. www.eu-robotics.net
2. Frey and Osborne, 2013, www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf
3. Bruegel, 2014, bruegel.org/2014/07/the-computerisation-of-european-jobs
4. www.rethinkrobotics.com/baxter
5. www.moley.com
6. www.hansonrobotics.com
7. www.willowgarage.com
8. h2020sarafun.eu
9. www.fp7rosetta.org
10. versatile-project.eu
11. robdream.eu
12. cordis.europa.eu/project/rcn/101283_en.html
13. cordis.europa.eu/project/rcn/100800_en.html
14. cordis.europa.eu/project/rcn/198763_en.html
15. cordis.europa.eu/project/rcn/106994_en.html
16. cordis.europa.eu/project/rcn/93725_en.html
17. robotics.eecs.berkeley.edu
18. cs.stanford.edu/group/manips
19. robotics.mit.edu
20. www.dlr.de/rmc/rm/en/desktopdefault.aspx/tabid-8016
21. www.is.mpg.de/en



Factories of smart autonomous equipment

Anders Helgesson, ABB, anders.helgeson@ch.abb.com
Mikael Rudin, ABB, mikael.rudin@se.abb.com

OVERVIEW

The overall objective of the improvements in the factory of the future is to improve cooperation between people, systems and equipment and, by this, to reach new levels of productivity.

In the future, the degree of automation in factories will increase. Future trends foresee that factories will be running high-volume and high-mix production flows, resulting in the need for more advanced and flexible automation solutions. This will also require more advanced and autonomous material logistics inside the factories. The dimension of digitalization will make it possible to monitor and also optimize the factory of the future resulting in shorter lead times, better quality and also lower production cost.

Market competitive forces will drive machine and equipment suppliers towards the delivery of smarter equipment and equipment as a service. This will lead to a situation where equipment has more functionality and larger parts of the functions will be provided from the cloud. The delivery of equipment as a 24-7-365 service will result in product-fleet-management centers supervising a product fleet. The evolution of plant equipment will go from monitoring through controlling and optimizing towards autonomous equipment; by analyzing collected factory data it is possible to feed back ”conclusion” to optimize/improve operations and processes. This will put a lot of pressure on equipment suppliers to create the required infrastructure.

The infrastructure and the ease with which new functionality can be added to an equipment will become a main competitive advantage for equipment suppliers. The capability for a factory to utilize this new kind of smart equipment will be an important factor in the productivity improvement.

INTERNATIONAL MATURITY

Globally, there are many initiatives looking at transforming the manufacturing industry. Most prominent in Europe would be Industry 4.0 in Germany but there are also initiatives in other parts of the world, such as China’s Made in China 2025 and India’s Make in India, all aiming at improving industrial productivity.

Historically, the automotive industry has been an early adopter of industrial robotics and has been driving implementation and use of flexible automation. Today, we clearly see other industries emerging, like for instance computing, communication, consumer electronics (3C), which puts even higher demands on automation flexibility due to the shorter product lifetimes in these areas.

The ability to use and create a collaborative environment for people, software systems and equipment has the potential to become a clear differentiator, for countries, in the strive towards new productivity levels.

CHALLENGES

From a technical point of view, there is always a trade-off between flexibility of the automated solution and the built in robustness of the solution. On one hand, you want a very flexible factory solution; on the other hand, you want high productivity without any loss in productivity or downtime (reliability) which is demanding for automation. This becomes even more difficult when the factories of the future will run high-volume and high-mix flows.

Another very important challenge is ease-of-use of the automated solutions, becoming more important as lead times are getting shorter and shorter. Ease-of-use is mostly targeting operators and engineering/commissioning phases, typically human-machine interaction but also machine-human interaction.

Specifically, assembly applications found in many industries provide a great challenge for industrial robotics especially from a robustness point of view, since assembly tasks are in general difficult for today’s robots in comparison with more common and simpler material-handling tasks.

Delivering new levels of productivity by the use of 10 000+ pieces of smart equipment, each with local and cloud-based intelligence, will present new challenges to the creation of large secure software systems.

TYPICAL INHERENT TECHNOLOGIES

The following are the main technology areas driving industrial robotics forward:

- robust and adaptive (intelligent) industrial robotics solutions, where more and more functionality is moved into smart software based on sensor inputs;
- ease-of-use of the total industrial robotics application
- safety solutions enabling machine-human collaboration
- design-to-cost and quality
- mega-factory manufacturing concepts, where scale lowers cost, still with improved quality;
- software architectures for systems of systems on a massive scale.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

There are many bottlenecks connected to the list above. However, at the right relationship between flexibility/robustness versus cost/quality in the automation solutions, this will drive a wider scale of implementation/adoption in industry.

Currently, there are substantial developments and investments in cost-competitive sensors and GPUs for intelligent robotics solutions, typically driven by autonomous transport solutions. Safety certification of many sensor combinations and software solutions are also emerging.

Software technologies like for instance apps, micro services and containers, have the potential to revolutionize the software landscape for equipment similarly to what happened for the mobile phones when they transformed to smart phones.

Software architectures spanning from embedded to cloud will be important for the delivery of the factory of the future built using smart equipment.

POTENTIAL FOR SWEDISH POSITIONING

Sweden, with its strong base in automotive, mechanical and electrical industries with global manufacturing footprints, should be able to leverage future factory concepts as mentioned above, with a strong focus on software and IoT development together with superiority in factory robotization and digitalization.

There is a lot of new ground to be broken in the software area for an industry built with smart equipment. There is certainly potential for several new large software players on the market in this area; Sweden has a lot of experience in many related areas.

POTENTIAL PLAYERS

Companies focusing on software for equipment and that has experience of building large software ecosystems using modern technologies like Apps, micro services, cloud, mobile and internet scale security.

NEED FOR SUPPORTING ACTIVITIES

If not already done, the Swedish innovation system would benefit from encouraging innovation and promoting start-up companies around the major technical universities in Sweden. This could be supported with accelerator programs or incubators (for instance the ABB initiative SynerLeap[1]).

To address the topics above, infrastructure for research must be modernized. Also, an environment with connection to venture capital and other global investors needs to be expanded.

The overall software skills in company management need to be improved to ensure that software is considered the main asset in equipment and systems in the factory of the future.

Footnotes and links:
1. www.synerleap.com

Part C:
Related trends and enablers

Distributed control

Anders Rantzer, LU, anders.rantzer@control.lth.se

OVERVIEW

Distributed control, in contrast to centralized control, means that sensing, computation and control actuation are carried out in many locations throughout a system, each with different available information and sometimes with different objectives.

Distributed control is necessary for operation of large infrastructures, such as networks for power, communications, traffic and water. It is also essential for operation of industrial production plants.

By definition, distributed control is a prerequisite for autonomous systems, since every kind of centralized control would make the agents non-autonomous. Moreover, distributed control is ubiquitous in living organisms, which creates interesting challenges and opportunities in healthcare as well as for environmental protection.

Development of new theory, methodology and tools for distributed control is critical to address tomorrow's challenges in infrastructure driven by urbanization, population growth and sustainable production. The economical drivers are very strong. New technology for sensing, communication, computation and actuation continuously creates new opportunities to lower costs and improve efficiency in the operation of large-scale systems through distributed control.

INTERNATIONAL MATURITY

The technology for distributed control is still rather immature. New approaches are developed separately in every application field, such as power systems, automotive industry, air traffic, process industry and health care. There are several reasons why the approaches differ between application areas, between countries and sometimes even within countries.

Priorities and economic incentives, but also legislation, are varying. Moreover, hardware investments for large infrastructures are huge, which means that decisions of the past have put severe constraints on future solutions.

Nevertheless, there are numerous examples of technology transfer between different application areas. For example, SCADA (Supervisory Control And Data Acquisition) is a control architecture which for several decades has been used in both in process industry and in power systems, as well as in gas pipelines and heating-and-ventilation systems.

The strategic importance of distributed-control technology is widely recognized internationally. For example, MIT recently created IDSS (Institute of Data, Systems and Society) across five MIT schools, with distributed-control experts in leading positions. Their web page says: "Our ability to understand data and develop models across complex, interconnected systems is at the core of our ability to uncover new insights and solutions."

CHALLENGES

Existing technology for distributed control relies to a large extent on the paradigm that different units should be kept sufficiently far apart, not to disturb each other. When studying traffic flow, this can be translated into safety distances between cars, trains and airplanes. For manufacturing plants, it means that there exists extra storage capacity between different units in a production line. For power networks, it means that there should be enough capacity of hydro-, nuclear- and coal power, to cover up for variations in wind- and solar power without interfering with consumers.

In all these areas, the separation paradigm is now being challenged by efforts to improve efficiency. Unfortunately, the lack of a systematic theory and methodology for distributed control is a bottleneck the transition towards more efficient and sustainable use of our resources. In many applications, this has resulted in debates between proponents of centralized and decentralized approaches, rather than a balanced distributed-control solution somewhere in between.

TYPICAL INHERENT TECHNOLOGIES

Distributed control has been one of the most important research directions in control engineering during the last 10–20 years. Significant progress has been made on fundamental aspects such as the interdependence between communication and control.

An important conclusion is that design and verification simplifies a lot if information travels faster through communication links than through the plant. Another important discovery is the derivation of design principles ensuring that dynamic effects cannot multiply faster than the static effects as they propagate through the system. Proper implementation of such design principles would have the potential to generate drastic efficiency benefits compared to earlier methods and penetrate into all major application areas.

Systematic implementation of distributed control also relies on enabling technologies in several disciplines, such as statistics, data science, information theory and inference, systems and control theory, optimization, economics, human and social behavior, and network science.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

The interdependence with other disciplines also means that distributed-control technology would greatly benefit from breakthroughs in these areas:

- Firstly, software for distributed control needs to be verified and validated in combination with dynamic models of communication devices and the physical environment. This requires new methods that combine control theory with computer science.
- Secondly, human-machine interaction is significantly more complex in a distributed setting than in a centralized one, so deeper understanding of the human

capacity of attention will be important.
- Thirdly, efficient control requires dynamic models. Automatic methods for maintenance of such models through learning and adaptation will be essential.
- Fourthly, many distributed control applications rely on technology for cyber-security to protect network the integrity. Lack of progress in this area could jeopardize future development of distributed-control technology.

POTENTIAL FOR SWEDISH POSITIONING

Sweden is very well positioned for taking a leading role in the development of distributed control technology, both on the academic side, where Sweden has a long and successful tradition of leadership in control theory, and on the industrial side, where the ability is high to take advantage of new developments. An important reason for this is that the PhD education in relevant subjects has been strong in international comparison during several decades. As a result, there are many PhD graduates in industry ready to take advantage of new academic research results.

POTENTIAL PLAYERS

For example, Swedish automotive companies are actively pursuing technology for autonomous driving as a means to improve traffic flow and reduce emissions. These efforts will rely on successful development of distributed control technology.

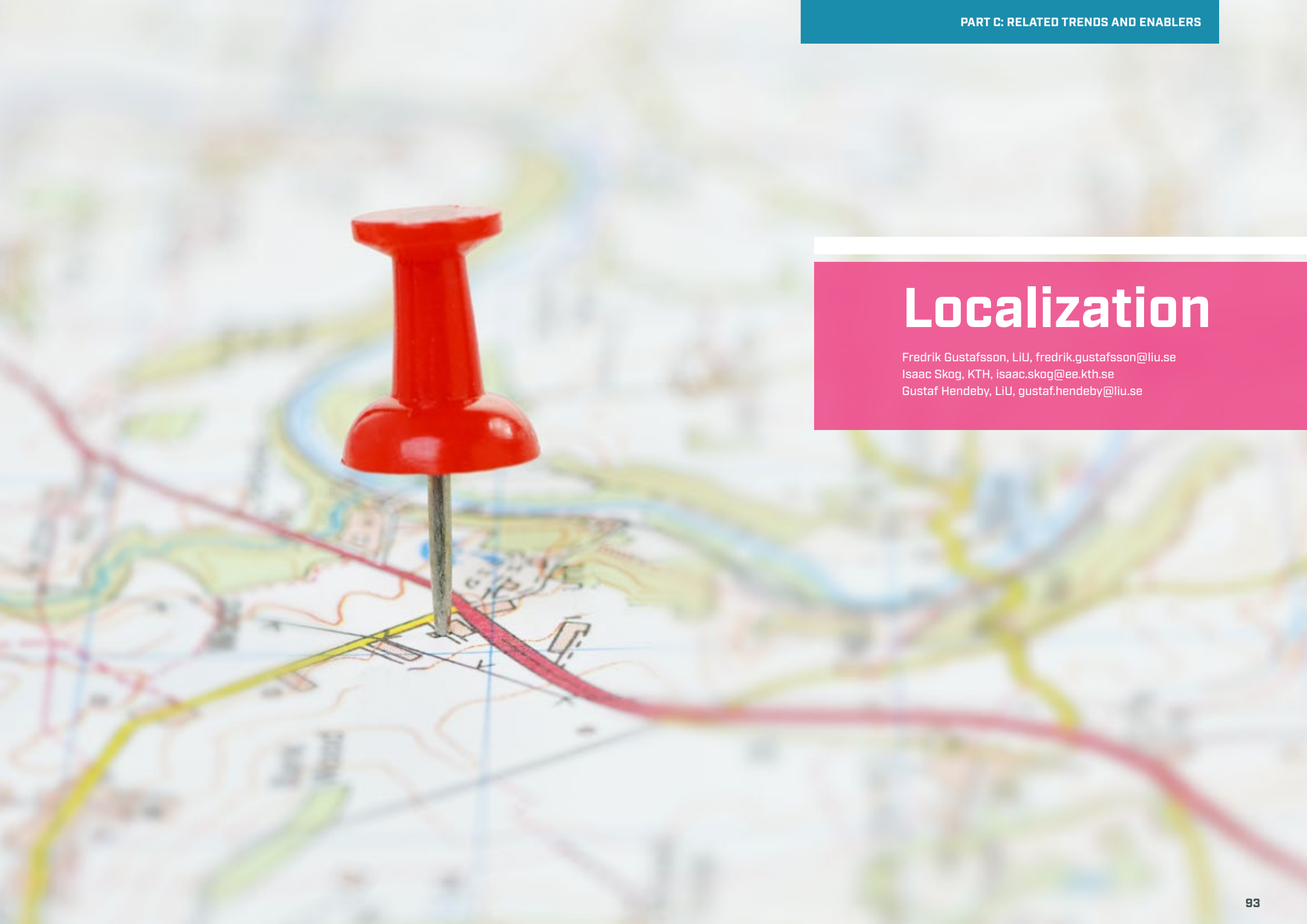
Another example is Ericsson, whose cloud- and 5G- products will be essential for implementation of new distributed control schemes.

Yet another example is ABB, with products essential for distributed control of industrial production and energy networks.

However, Sweden also has a track record of smaller companies producing software and consulting services for control and optimization of advanced processes. A good example is the group of companies making software and services for modeling and simulation in Modelica.

NEED FOR SUPPORTING ACTIVITIES

Prototype implementations of large-scale networks in many application areas are currently being in built all over the world. However, it is often hard to get access to relevant data for comparisons and academic research. Systematic development of open databases with dynamic data for major applications in a Swedish context, would be extremely valuable to support Swedish research and innovations in this field.



Localization

Fredrik Gustafsson, LiU, fredrik.gustafsson@liu.se
Isaac Skog, KTH, isaac.skog@ee.kth.se
Gustaf Hendeby, LiU, gustaf.hendeby@liu.se

DEFINITIONS

Localization: The process of determining an object's location relative to a predefined coordinate system.

GNSS: Global Navigation Satellite System.

INS: Inertial navigation system.

MEMS: Micro-electro-mechanical systems.

SLAM: Simultaneous localization and mapping.

OVERVIEW

The words localization and positioning are synonyms, and refer to the process of determining an object's location relative to a predefined coordinates system. Navigation is the task of monitoring and controlling the movement of an object from one place to another, and localization is therefore a fundamental part of the navigation process.

Maps play an important role both in the positioning and navigation process as they provide information about the feasible positions of the object and how to move the object from one point to another. SLAM (simultaneous localization and mapping) denotes the combined task of localization and creating (or updating) the map on the fly.

Target or object tracking is a closely related area, originating from military applications. The term tracking is often used when a network performs the localization of an object, rather than the object itself. However, when everything is connected, the difference of tracking and localization disappears.

Situational awareness concerns localization in a local context, where the relative position to stationary and moving objects is crucial. Here perception is an often used term, where computer vision is an important complement to pure localization technologies.

Accurate localization anywhere and anytime – of vehicles, robots, humans and gadgets in both the absolute and relative sense – is a fundamental component needed in the development of intelligent and autonomous products and systems. Localization has been an enabler for all kind of transportation systems, from the ancient seamen exploring the globe and making the first world maps, to modern aircraft that can fly and land autonomously. Localization of vehicles is of industrial relevance for many Swedish companies such as Volvo Cars, Volvo Trucks and Scania, which are developing autonomous road vehicles, as well as Atlas Copco and Volvo CE, which are developing solutions for smart mining.

Moreover, as localization is a fundamental capability in the design of smart products and systems, it is an enabling technology for hospitals, industry asset management, infrastructure for traffic, logistics, water, gas, electricity and building control and similar. Localization is also central for certain applications in the gaming industry, and is essential for many AR and VR applications.

INTERNATIONAL MATURITY

Today, there is no super-sensor or generically applicable technology that can provide anywhere-and-anytime localization, and companies are applying the technology that is best suited to their needs. The most frequently employed localization technologies are discussed below.

Satellite-based localisation

Nowadays, satellite-based localization using the different Global Navigation Satellite Systems (GNSSs) is the most commonly used technology for localization in outdoor transportation systems, including manned and unmanned aircraft, surface vessels and road vehicles. GPS appeared in the early 1990s as the first GNSS, and today there are also a handful of other systems in use or in deployment.

Generally, a standard GNSS receiver provides the position with a ten-meter accuracy, within 30 seconds from when it is turned on. With ground base stations and phase measurements, the accuracy is improved three orders of magnitude. The assisted GNSS-receiver technology used in smartphones has a time to fix that is about one order of magnitude shorter. The power consumption is also one order of magnitude smaller.

Cloud-assisted GNSS-based positioning has the potential to bring down the power consumption another three orders of magnitudes. This is important in certain IoT applications, but the technique yet remains largely unexplored and there is a great potential for technological leaps within the area.

Tight integration of GNSS into navigation systems is based on the so-called pseudo-range to each satellite, in contrast to loose integration that uses the GNSS position in a systems-of-systems approach, and tight integration that enables both improved performance and robustness. The new generation of Android enables access to the satellite pseudo-ranges, which opens up for many new sensor-fusion algorithms exploring redundancy with other information sources.

Inertial navigation

Before the birth of the GNSS, inertial navigation was the primary technology used in high-end localization systems. Inertial Navigation Systems (INSs) are still used in aircraft and surface vessels, but are also crucial for underwater vessels, rockets and other vehicles operating out of range of GNSS signals.

An INS can provide relative position information of high accuracy for a long time, but the inherent drift due to accumulated sensor errors will sooner or later limit its usefulness. However, high-grade INSs are still expensive, and

unfortunately no technology breakthrough that will drop the prices significantly in the near future is foresighted.

An INS consists of an inertial measurement unit (IMU) and software. Today, IMUs are manufactured in large volumes in MEMS technology, which provides the measurements needed in INS in our smartphones and other gadgets. The performance, price and power consumption have all improved enormously over the past ten years, basically since the first iPhone appeared. However, modern road vehicles do not include a full MEMS suite for INS, but rather a subset of accelerometers and gyroscopes for certain functions (yaw control, airbag, roll-over detection and similar).

Radio-based localization

Radio localization is an old technology. The first application was for sea navigation, where the lighthouses were complemented with a rotating radio beam used for bearing computation not requiring visibility. Several other radio-based systems have been developed for sea navigation, but are not in use today due to the superior accuracy provided by the GNSSs. Dedicated radio-localization systems are used in other applications as well. For instance, airfields have local networks of radio beacons for positioning approaching aircraft.

Currently, cellular networks are driving the development of radio-based localization technology. Location-based services appeared already in the GSM system, where the requirement to locate emergency calls with an accuracy of some 50 meters was a driving factor. Rather surprisingly, the basic measurements for localization have remained the same in 3G and 4G systems. However, for the upcoming 5G system, the standardization work includes localization aspects, including new type of measurements and network assistance technologies.

Wi-Fi networks are frequently used for localization in GNSS-denied environments, for instance in indoor environments such as factories and mines. There are no large-scale deployments or standard solutions for Wi-Fi localization yet on the market, but there are hundreds of companies active in this area. Demonstrated accuracy is in the order of meters in areas with good Wi-Fi coverage.

Like cellular networks, Wi-Fi networks were designed primarily for communication, not for localization, but new hardware for the access point will enable dedicated measurements for localization, such as distance (time of flight) and bearing (directional antennas).

Bluetooth Low Energy (BLE) technology introduced in Bluetooth 4 standard is another enabler for localization. BLE tags are small and energy efficient, and with a regular-sized battery the lifetime can be hundreds of years. This provides a huge potential for many IoT applications, where the position and ID of the devices can be automatically determined by the infrastructure, possibly with some ten-meter accuracy.

Feature-based localization

Feature-based localization, where observed features are correlated with the information in a map, have been the main tool for localization in all times. Traditionally, it has been done by a human matching visual observation to the features described in the map. Today, the same basic principle is commonly employed by missiles, aircraft and submarines, where height (depth) measurements are correlated with a topographic map; a process referred to as terrain navigation.

Lately, new kinds of maps are emerging, in particular maps of radio-signal strength from various radio transmitters and maps of the disturbances in the local magnet field. A recent trend is to look at how crowdsourced radio-signal and magnetic-field maps can be used for cloud-based localization.

CHALLENGES

Thanks to its global coverage and scalability, the GNSS technology has become the de-facto standard for outdoor localization. However, for localization in urban areas, indoor environments and underground, as well as in safety-critical applications, the GNSS technology cannot deliver the accuracy, reliability and coverage needed. Current challenges therefore focus on developing cost-efficient localization solutions for environments and situations where GNSS-based localization solutions are not sufficient. Future enabling localization solutions should fulfil the following three performance metrics:

1. Scalability over geographical area and cost of deployment.
2. Availability over space and time (anywhere-and-anytime localization), including an ability to adapt over space and time.
3. Reliable, providing trustworthy position information, not only concerning the position accuracy, but also an uncertainty description.

The main approaches at hand are based on sensor fusion and learning, where redundancy and large data sets are tools to achieve the goals. Cloud solutions including crowdsourcing concepts provide promising ways to access large amount of data from various sensor modalities. New sensor technologies and radio standards offer a steady stream of new sensor fusion challenges, and old paradigms have to be revised for each technology leap. Centralized off-line sensor fusion solutions provide upper bounds on performance, while distributed real-time implementations on incomplete data streams pose the practical challenges.

All in all, the challenge is to research and develop scalable, adaptable and reliable localization algorithms that can – using sensor information fusion and collaboration between nodes and user – learn, adapt and optimize their behavior to the usage conditions. Development of hierarchical data-processing strategies that merge cloud-based and distributed localization methods will be a cornerstone to ensure both scalability and reliability of the systems.

Today, the only type of system that meets the anywhere-and-anytime requirements is high-grade INS, an area which is more or less dead today in research and development, and which is certainly not a scalable solution due to the high cost. Most current development and

research efforts are focused on infotainment (as location-based services) or human navigation support. There are a few exceptions, all based on the principles of information redundancy and sensor fusion. A few examples are:

- Google Location Service appeared on smartphones some five years ago as a solution based on redundancy. When GNSS-signal reception is not possible, it looks for nearby Wi-Fi networks, and if any of them appears in Google's location database. If not, it uses the cellular network information, which has an almost global coverage, and again looks up the most plausible position in another Google database.
- The GNSS-receiver support/backup/replacement system for cars developed by NIRA Dynamics 2002 uses GSM for finding the relevant cell, then implements a kind of INS solution just using the wheel speeds, and compares the trajectory with a street map. In this way, a demonstrated accuracy on the meter level is achieved.
- The indoor positioning system (IPS) developed by Senion, utilizing all sensing modalities in a modern smartphone. The IPS algorithm switches from GNSS localization outdoors to a sensor-fusion algorithm based on a combination of Wi-Fi and BLE measurements together with INS and building-map information, where SLAM methods are used to create and update a fingerprint map. That is, more or less all basic principles are used and the result is meter accuracy in covered areas.

TYPICAL INHERENT TECHNOLOGIES

There are a few specific technologies standing out as particularly interesting:

Sensor fusion

Since there will never be a super-sensor for localization, all future development must be based on a variety of redundant information sources. Here, the sensor-fusion concept is fundamental: how to combine location-based measurements from different sensor modalities, sensed at different positions, with information from maps and other non-sensor-based information sources? Further, localization algorithms can be embedded into a filter by using prior information of how the object moves over time. The Bayesian paradigm provides a coherent framework for sensor fusion, and there is today a mature theory, where sample-based Monte Carlo methods are the most flexible ones. Here the particle filter is the state-of-the-art solution to many problems, but in its current shape it has scalability limitations.

Mapping and SLAM

Mapping is a crucial area for most future localization principles. For mapping, a ground truth of position is needed, and for this single purpose dedicated localization technologies and mapping tools can be used. The most important example today is how RTK GPS is used for getting centimeter accuracy of maps. For mapping indoor environments and mines, special equipment is also used.

Maps can be created on the fly in a SLAM framework, where the position is estimated jointly with the map. The theory emerged in robotics, where the main demonstrator was a wheeled robot moving in an office environment. This theory matured some ten years ago, but SLAM as a concept of creating maps without dedicated equipment is still a hot topic.

Crowdsourcing

However, to get a scalable solution and an adaptive map, crowdsourcing is the main tool. Crowdsourcing can be seen as a huge collaborative effort to attack the SLAM problem. Google used the Ingress app to map signal strengths in the cellular network, and further used the gaming to move users to places where the radio map needed to be updated. This radio map is today used in Google location services.

Tesla is claimed to crowdsource landmarks along the roads from the camera, radar and lidar to map the environment. In any smartphone today, both the phone manufacturer, chipset manufacturer and app providers log location-based data to their own proprietary databases.

Open databases

In the future, there is a need for open databases, and market places for location-based information. The initiative Here, to mention one promising example, is supported by several German OEMs and many other actors, and it is quite interesting in this context. It is based on the Nokia street map, and there is a business model for data providers, data processors and data consumers to exchange data and services in the cloud.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

It is of course impossible to foresee the future, in particular disruptive events that give leaps in developments. We here mention a few trends that can frame the developments in the next ten years.

New radio standards will include new dedicated location-based measurements. What is about to come is Wi-Fi access points with range (time of flight) and bearing (directional antennas) measurements. 5G will likewise provide similar measurements. Massive MIMO is not directly focused on localization, nor providing any location relevant measurements, but the huge number of sensor elements have the potential to make leaps in localization technology.

Legislations can influence the speed of technology development, for instance the upcoming US 911 positioning requirement indoors. One consequence of the requirement to locate cell phones to the correct floor is that this will force mobile phones to have a built-in barometer and report barometric pressure to the network. This also influences standardization, where measurements must be provided in new standards to meet the requirements. Why should not future automotive vehicles also have a legislated accuracy of their localization systems?

The ever-increasing number of GNSS systems and satellites in use will only marginally improve the situation, since the inherent problem is the low signal strength at earth (–140dB, which is 5–10 order of magnitudes weaker than other radio technologies). However, network-assisted GNSS solutions can bring down the energy requirements several orders of magnitude, which can be an enabler for battery-operated IoT applications. Infrastructure and cloud services are crucial for achieving leaps in localization performance. Open standards are lacking today, but would certainly open up for many new inventive players on the localization arena, not only the owners of the proprietary databases we have today.

To current date, the focus of the MEMS-sensor-technology development has been on reducing the size and power consumption of the sensors and measurement performance of the sensors has been secondary. However, recent market surveys predict that focus of the ultra-low-cost-sensor development now will turn towards improving the performance of the sensors to meet increasing demands of the automobile manufacturing industry. One potential way to this is to make arrays out of ultra-low-cost sensors and fuse their measurements to create virtual "super" sensors.

As localization and motion sensors become ubiquitous in all types of products and systems, the potential to crowd-source information for building of maps (radio-signal strength, magnetic-field disturbances and similar) will drastically increase. The possibility for development of collaborative localization strategies will also drastically increase. Fundamental for this is the development of hierarchical as well as distributed information fusion strategies; the hierarchical strategies are needed to not flood the networks with data, and the distributed strategies are needed to ensure robustness against changing network constellations and similar.

Potential for Swedish positioning

Sweden in general and WASP in particular have an excellent background in localization theory and applications. The strengths in Sweden lies in the software rather than the hardware. We are in the international frontier for radio-based localization, Bayesian and adaptive algorithms, and support systems for GNSS and INS, to mention a few theoretical areas. Public safety and security, including critical infrastructure protection, are other areas where Sweden has a strong standing.

Sweden is not participating in developing GNSS core algorithms, and we do not have many big players with a sensor focus relevant for localization.

POTENTIAL PLAYERS

Ericsson has the resources and competence to influence localization standardization and development for cellular systems.

Swedish industry has recently refocused resources to the autonomy area. Here we can mention localization efforts by Volvo Cars, Volvo Trucks, Scania, Autoliv, NIRA for road-bound vehicles and Volvo CE, Atlas Copco, BT/TMHE, Husqvarna for other wheeled vehicles.

Localization as an enabler for smart factories is explored by ABB, SKF, BT and similar.

Localization for smart mining includes Boliden, LKAB, Atlas Copco.

There are many startups in the gaming industry, for VR/AR applications and for indoor positioning systems that would benefit for progress in the localization area.

NEED FOR SUPPORTING ACTIVITIES

The theory and applications of localization are developing quickly, and there is a need for rapid prototyping solutions that can be used for evaluation of new concepts. For this, the most urgent need is a scalable and flexible infrastructure that enables experiments and field trials with a wide variety of sensors, information sources, and sensor fusion algorithms. We foresee a hardware and software infrastructure that is:

- **Mobile:** the infrastructure should be easily movable to enable evaluation of scalable solutions over space in places that fit the application, not the technology.
- **Multi-purpose:** the infrastructure should fit (include) a variety of sensor modalities and be applicable to public safety, autonomous driving and other application areas, and include all kind of communication interfaces (cellular radio, Wifi, Bluetooth, TCP, USB) to sensors and network.
- **Modular:** the infrastructure should be possible to combine in a sensor network and with other hardware (UAV, cars, sea vessels, robots and similar.) and software (cloud solutions, visualization interfaces and similar).

In its simplest form, the infrastructure can be conceptually realized using a set of smartphones as sensor-nodes, a local server as a backend, and specially designed software for high-level control of sensors, fusion algorithms, and communication protocols. Such a baseline infrastructure can be extended to be self-powered and cooperate with robust sensor nodes and cloud assisted computations.

Intelligent hardware and materials

Per Ingelhag, Ericsson, per.ingelhag@ericsson.com
Astrid Algaba Brazalez, Ericsson, astrid.algaba.brazalez@ericsson.com
Jonas Hansryd, Ericsson, jonas.hansryd@ericsson.com

OVERVIEW

Autonomy is the ability of a system to achieve goals while operating independently of external control. There is a spectrum of autonomy in a system that ranges from local autonomy within a subsystem, where actions may be executed in response to a stimuli or local information, to system-level autonomy, which manages actions and handles constraints across subsystems. Fully autonomous systems would be able to act independently and intelligently in dynamic and uncertain environments.

Autonomy's fundamental benefits are increasing system operations capability, enabling cost savings by reducing human labor needs and increasing efficiencies, and increasing robustness in uncertain/changing environments.

The need for autonomy is evident when:
autonomous decision making reduces overall cost or improves effectiveness;

- local decisions improve manageability and robustness of overall system architecture and reduce complexity;
- decision making is beyond communication constraints (delays and communication windows);
- decisions are better informed by the richness of local data compared to limited linked data;
- time-critical decisions must be made onboard a system or vehicle, such as control, health, and life-support.

A system, for example mobile radio system, is comprising hardware and software. A hardware subsystem itself could be autonomous and intelligent – intelligent hardware – if the autonomous system (AS) is local to a hardware block/function/structure or even material – intelligent material. A pre-requisite is that the intelligent hardware/material shall change its properties based on the current situation without outside intervention, except new high-level control. To do this, the intelligent hardware/material relies on:

- **sensing** – sensors sensing the condition, including physical quantities as well as internal data flow probes;
- **perception** – an autonomous-and-adaptive-system (AAS) algorithm running on an embedded compute platform for automated situation analysis and decision making;
- **manipulation/modification of the hardware** – flexible hardware structures or material with properties that could and is desirable to be changed in the application.

In principle, any hardware block/structure or material that is flexible and with physical quantities possible to observe and desirable to be controlled in a AS context, could qualify to a WASP research area. This is obviously a rather big scope. Therefore, this foresight on intelligent hardware and materials are mainly discussed in the form of "application areas" in the typical inherent technologies chapter, whereas some intelligent material based on meta-materials are more thoroughly covered as an example on emerging technologies through the foresight. Metamaterials are artificial materials that provide special properties that cannot be found in nature.

INTERNATIONAL MATURITY

The context of reconfigurable metamaterials is quite novel, and its popularity has increased during the last five years especially focusing on wireless communication applications. The development of the context so far depends pretty much on the operating frequency band since the key factor lies in the tunable components applied to the metasurface.

At lower microwave frequencies, varactor and PIN diodes have been used for electronic tuning[1,2]. Nevertheless, their use has severe constrains at higher millimeter-wave frequencies due to high loss, parasitic effects, and nonlinearities. At lower millimeter-wave frequencies, microelectromechanical systems (MEMS) have been successfully employed.

However, to obtain tuning, a MEMS component has to be integrated to each element of the periodic structure or metasurface. This implies an upper frequency limit for the suitability of the technology, as the dimensions of the periodic elements decrease with the frequency, not allowing for the incorporation of an electrically large component in the unit cell of the hypothetical antenna array.

Also, tunable materials such as ferroelectric substrates at lower microwave frequencies and more recently liquid crystals at higher millimeter-wave frequencies have been investigated producing promising results. The main disadvantage of these tuning techniques is that they exhibit high losses and very low switching speeds (in the case of liquid crystals). Recently, piezoelectric materials and actuators have been also proposed for the dynamic reconfiguration of phase-shifting surfaces[3].

Due to the novelty of the topic we are not sure we can define consolidate international leading players of the context. There are relevant world-class academic groups that have been working in this area and presenting interesting results and ideas at international conferences and symposiums like the annual European Conference on Antennas and Propagation (EuCAP) or the European Microwave Week. We have gathered some names of researchers whose groups have been contributing from the academic point of view around reconfigurable metamaterials:

- Prof. Anthony Gbric, University of Michigan, USA[4]
- Prof. Alexandros Feresidis[5], University of Birmingham, UK[6]
- Prof. Stefano Maci, University of Siena, Italy[7]
- Prof. Oscar Quevedo Teruel, Kungliga Tekniska högskolan (KTH), Sweden[8]

About industry commercializing reconfigurable metasurfaces focused on antenna applications, we can remark Plasma Antennas[9]. The applied technology is described in upcoming sections.

CHALLENGES

Some challenges in the area of reconfigurable metamaterials are the following:

High-frequency issues

The need of higher data rates in mobile communications implies a move upwards in frequency (there is a high interest in the millimeter-wave frequency range) which leads

to different kinds of challenges regarding radio-frequency (RF) component design and antenna integration. At high frequencies, the way of designing components changes since we need to provide high integration meaning suitable component interconnection together with the antenna; loss becomes a high issue if using conventional materials.

Active components

There are components that are very relevant for providing reconfigurability features to metasurface antennas such as PIN diodes, transistors, varactors, RF-MEMS to switch between frequency bands and modes. Research of this type of active components that can perform well at the high range of millimeter-wave frequencies is critical.

Tolerance sensitivity

Metasurface antennas offer potential to be applied at millimeter-wave systems, but as frequency increases, manufacturing and assembly tolerances become more critical and affect the overall performance. Some recently reported metasurface antennas have shown that some of its design parameters are very sensitive to tolerances, and thereby, a tolerance investigation is essential within the design process of reconfigurable metasurface antennas or components in order to ensure a reliable functionality of the autonomous system.

TYPICAL INHERENT TECHNOLOGIES

Research areas could be in any of the general parts of an ASS: sensing (hardware sensors), perception/analysis/decision making (efficient compute platforms) and the hardware block/structure/material itself to be modified/controlled.

Research areas of interest are areas where design methods, hardware and materials are enhanced using an ASS approach (enhanced in this context meaning lower cost, higher performance, less power, higher reliability, less complexity and similar).

Autonomous systems application view

Application – hardware systems

- Self-adaptability/management of context-aware hardware sub system. Typical application is hardware self-calibration and configuration:
  - Digitally assisted calibration/tuning of analog signal chains – from data converters to low-noise amplifiers, power amplifiers, including antenna and other passive networks (tunable filters). Typical applications with varying maturity are data converters.
  - Self-configuration of signal-chain parameter based on context.
  - Example of principle applied on ADC: "A 13b 4GS/s Digitally Assisted Dynamic 3-Stage Asynchronous Pipelined-SAR ADC" by Xilinx, presented at ISCAS 2017[10].

- Hardware sub-system autonomous anomaly/fault detection and management:
  - Monitors, predicts, detects and diagnoses faults and accommodates or mitigates the effects (repair using redundancy, resource reallocation for graceful degradation).
  - Hardware ageing (lifetime monitors) management for optimal in-context reliability.
  - Prognostics and health management (PHM), enables the determination of when scheduled maintenance makes sense, minimizing system life-cycle costs.

- Intelligent radio base-station energy systems based on machine intelligence (MI):
  - Self-learning power management and optimal operation point (frequency, supply voltage, back bias) based on context (process voltage temperature/history and ageing/prediction and load/load step).
  - Optimizing for cost and availability (including performance degradation by load throttling) for in-context sun, wind, grid, diesel power.
  - Algorithms and cost functions for power sources.

Application – models and methodology

- Verification and validation of ASS hardware:
  - Efficient tools and techniques for verification and validation (V&V).
  - Performance evaluation of self-adaptable context-aware systems.

- Modeling and simulation of ASS hardware:
  - Domain-specific modeling and simulation of sensing and hardware tuning.
  - Models of self-adaptation hardware mechanisms.

- Design methods for self-adaptable context-aware systems:
  - Approaches to model and represent context adaptability, self-adaptability and self-manageability.
  - Architectures and middleware models for self-adaptable context-aware hardware systems.
  - Approaches to the feedback of autonomous system data statistics back to the design phase – statistical design methodology based on mission data.
  - Approaches to the feedback of autonomous system data statistics back to manufacturing adjustments.

Application – hardware (sensor, compute, material structures)

- Efficient CMOS compute platforms for autonomic perception:
  - Cost- and power-efficient compute resources/platform for AS algorithm. Application specific.
  - Integration solution/materials for 3D integration in package to mitigate Moore's-law slow-down.
  - Affordable cooling solution for 3D integration. Can new materials (graphene films) or structures integrated vacuum chambers be used? How to spread/dissipate power form a high-output-power array-antenna system for above 25 GHz (small due to wavelength). Water cooling?
  - Cost-efficient material/solutions/semiconductors that can withstand high temperature with maintained

performance and reliability. This is also applicable for power amplifiers (meaning GaN).

- Redundancy concepts:
  - In compute platform.
  - In signal chain.

- Cost-efficient in-system signaling:
  - Silicon photonic integration, embedded wave guides – or other distribution – would move the architecture limitations from physically/electrically-restricted topology to a more free "logical topology" supported by fiber/silicon photonics.
  - Fiber connector, optical waveguides and optical cross connect. Manufacturability/Assembly

- Post-CMOS era computing:
  - For example, quantum computer.
  - Neural networks hardware.

- Connectivity building practice – cost-efficient 3D antenna-array integration of TRX and antenna, including filter and front-end:
  - Intelligent materials.
  - MEMS.

- Sensors:
  - Sensor fusion.
  - High-resolution-range sensors
  - Invitro sensors.

Intelligent materials/reconfigurable metamaterials

Metamaterials are artificial materials that provide special properties that cannot be found in nature.

They are usually arranged in periodical patterns at scales smaller than the wavelengths of the phenomena they influence. Metamaterials derive their properties not from the properties of the base materials, but from the entire structure as a whole. In this way, we can achieve benefits that go beyond what is possible by using conventional materials.

Metamaterials as a concept has recently seen a tremendous growth in the research community in fields like acoustics, photonics and radio-frequency (RF) engineering, and in particular antenna applications.

Metasurfaces are the 2D equivalent of metamaterials, and they are able to modulate the behavior of electromagnetic waves through a specific boundary condition. Metasurfaces are employed to allow or prohibit the electromagnetic wave propagation in certain frequency bands and directions.

Metasurfaces can be used as platform for exploration and modelling of new physical effects that can be interesting for autonomous systems and the development of practical sensor solutions.

These materials have been widely applied into antenna systems and microwave/millimeter-wave passive-component design for solving problems that exist when using conventional materials (mutual coupling, high loss, leakage, signal interferences and similar). During the last five years, it has spanned the applications from reconfigurable antennas for multiple-input-multiple-output (MIMO) applications to autonomous arrays and printed RF-MEMS switches.

Reconfigurable antennas that can adapt to the needs of the environment or different user cases is an interesting research topic. Antennas of this type may constitute a relevant hardware part of an autonomous system where continuous connectivity is a key factor.

Frequency tuning, polarization adaptability, radiation pattern shape reconfigurability, autonomous switchable beams and electronically scanning antennas without the use of phase shifters are some of the interesting features that a reconfigurable metasurface antenna system could provide.

EXAMPLE: LENSES

An example of self-switchable beam (or selectable multibeam) lens antenna that has been recently introduced is the so-called plasma antennas. This type of antennas is commercialized by Plasma Antennas[3].

This could be a candidate to be used as self-reconfigurable multibeam antenna on an autonomous system where a novel microwave-optic technology that exploits reflective properties of solid-state plasma created in an array of PIN-diodes is applied.

Their products consist of a plasma-silicon device (PSiD) providing fast, electronic beam forming and beam selection functions. A PSiD can be regarded as a multi-port, wideband, commutating switch that replaces high-loss RF switches, phase shifters and attenuators with one compact device.

Due to their silicon integrated-circuit construction, PSiDs can be reproduced to high precision for the mass market at low cost. Moreover, they exhibit high power handling and, unlike RF MEMS, can be "hot switched" (with power on).

Technically speaking, plasma materials are not usually classified as metamaterials but constitute a type of artificial material and sometimes included into the metamaterial family.

Plasma Antennas are currently developing 28 GHz and 60 GHz devices for integration into point-to-multipoint backhaul as well as wireless-gigabit and 5G access points. They are also working on a 77 GHz device for automotive scanning radar for autonomous vehicles.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

Successful application of intelligent hardware and materials will generally lead to increased system-operation capability, enabling cost savings by reducing human labor needs and increasing efficiencies, and increasing robustness in uncertain/changing environments.

New research breakthroughs in any of the three areas (sensing, perception, manipulation/modification of hardware) constituting an AAS is important. Perception is likely to at least to some part rely on machine intelligence and/or big-data analytics. Any breakthrough in these areas relevant for AAS can be important.

POTENTIAL FOR SWEDISH POSITIONING

The use of intelligent hardware and artificial materials that are self-tunable and can adapt to different environments or user cases could find good possibilities in automotive, aircraft and mining industries developing both radars and hardware systems involved in autonomous vehicles.

Sweden has many players working in this category that could have a relevant role as industry research partners, like ABB, Atlas Copco, Boliden, Volvo, Zenuity, Sandvik, Scania and Saab in collaboration with universities and research institutes like Chalmers, KTH or SP.

Power and cost-efficient systems are essential for allowing continuous connectivity between devices which is a key factor in an autonomous system. Ericsson is working with many partners and has a profound Swedish background in terms of connectivity, and can therefore play a role that will allow the technology to develop faster compared to in other places. With the same reasoning, aside from connectivity, a system that enables massive-scale device life-cycle management is needed.

POTENTIAL PLAYERS

The area of intelligent hardware and materials is applicable to most WASP partners.

Regarding metamaterials, KTH has a newly established research group working in the field of high symmetric metasurfaces whose project leader is Professor Oscar Quevedo Teruel.

Microwave Electronics Lab of the Department of Microtechnology and Nanoscience at Chalmers is a world leading research group working with THz electronics including sensor systems. There are different small start-ups working in active components for millimeter-wave applications that could provide a great input into this research collaboration.

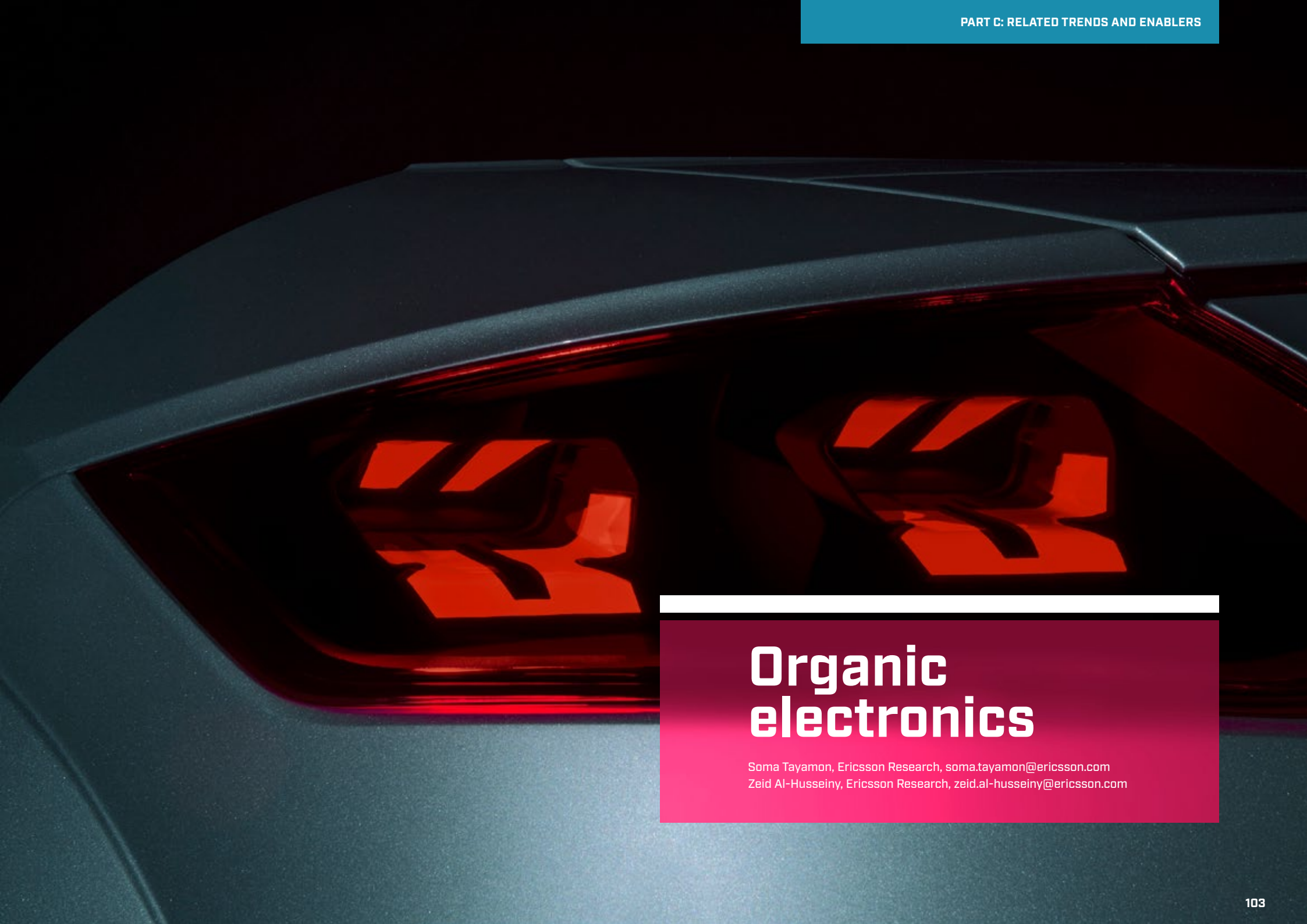
NEED FOR SUPPORTING ACTIVITIES

Research collaborations between industry partners and academia are essential to developing new technologies. Current players in Sweden – companies, research institutes and universities – have all excelled in their different research fields. However, for making intelligent hardware and materials applied to autonomous systems to become an area with consolidated potential, the research collaborations between the different Swedish players, both from industry and academia, need to have a better bridge and find advantage of a multidisciplinary environment to achieve the wanted project goals.

Funding and mobility programs could be beneficial to achieve better synergies. For example, with mobility programs, grants researchers from universities can visit industry for a period and do a research stage in order to exchange knowledge in both ways.

Footnotes and links:

1. S.V. Humand, J. Perruisseau-Carrier, “Reconfigurable reflectarrays and array lenses for dynamic antenna beam control: A review,”IEEE Trans. Antennas Propag., vol. 62, no. 1, pp. 183–198, Jan. 2014.
2. R. Guzmán-Quirós, J.-L. Gómez-Tornero, A. R. Weily, and Y. J.Guo, “Electronically steerable 1D Fabry-Perot leaky-wave antenna employing a tunable high impedance surface,” IEEE Trans. AntennasPropag., vol.60, no. 11, pp. 5046–5055, Nov. 2012.
3. M. Mavridou, A. P. Feresidis, P. Gardner, and P. S. Hall, “Tunable millimetre-wave phase shifting surfaces using piezoelectric actuators,” Microw., Antennas Propag., vol. 8, no. 11, pp. 829–834, Aug. 2014.
4. web.eecs.umich.edu/faculty/grbic
5. M. Mavridou, K. Konstantinidis, A. P. Feresidis “Continuously Tunable mm-Wave High Impedance Surface,” IEEE Antennas and Wireless Propagation letters, vol. 15, 2016.
6. www.birmingham.ac.uk/staff/profiles/eese/feresidis-alexandros.aspx
7. www.dii.unisi.it/~macis
8. www.etk.ee.kth.se/personal/oscarqt
9. www.plasmaantennas.com
10. www.xilinx.com/support/documentation/product-briefs/rfsoc-ieee-paper.pdf



Organic electronics

Soma Tayamon, Ericsson Research, soma.tayamon@ericsson.com
Zeid Al-Husseiny, Ericsson Research, zeid.al-husseiny@ericsson.com

DEFINITIONS

LED: Light-emitting diode.

OLED: Organic light-emitting diode (or organic LED).

E-Textile: Electronic textile (a.k.a. smart textile).

IoT: Internet of things.

OPD: Organic photodiode.

CF: Cystic fibrosis.

UTI: Urinary-tract infections.

GABA: Gamma-aminobutyric acid.

OEIP: Organic electronic ionic pump.

E-Skin: Electronic skin.

E-plants: Electronic plants.

OVERVIEW

The concept of intelligent hardware and materials covers a vast area of technologies used in devices and products. As an example, materials that change shape, colour or properties during exposure to light, heat or pressure, are considered intelligent materials. Such materials can be used to create hardware that can make autonomous decisions without human interaction; this hardware is considered intelligent hardware. In this text, we focus on organic electronics, which is when the electrolyte in the electronics of the device consists of organic components.

Introducing organic materials into electronical devices removes the barrier between biology and technology. These devices may be used as a translator between the biological and technical world. If a biological signal can be measured and translated into an electronic one and vice versa, the possibility of manipulation and regulation of biological and physiological systems is created.

One important property of organic electronic devices is that due to different organic materials that can be used in the device, different properties on the device are achievable. These properties include not only physical properties such as shape, flexibility and elasticity of the material, but also chemical properties.

Using organic electronics, biomedical devices can be integrated into the body, without being repelled or causing any harm. Wearables and/or augmentation can be placed directly on, or integrated, into the skin as well. Organic electronics enable printed electronics, easily mass manufactured at a low cost, which in turn enables cheap disposable electronics which creates new possibilities in all industries.

Intelligent hardware, in the form of organic electronics, allows for optimizing performance through adaptation to the environment using software closely coupled with the hardware.

INTERNATIONAL MATURITY

In the context of research within organic electronics, the leading players are universities. Some start-up companies have started production of different devices and products in different fields such as printed electronics[1,2], and wearable temperature stickers[3].

The concept of organic electronics is subject of a continuously increasing interest. Technology fields within this context are rapidly growing, both in broadness but also in depth. Research areas such as e-plants and e-skin (see below) are receiving more attention as the potential in such fields is becoming more recognized. Furthermore, there is a significant focus particularly revolving around research on organic devices for medical usage, where the fruit of such research could potentially move the medical industry into a whole new age.

The topics in which major companies have established some ground include OLED (organic light-emitting diode or organic LED) where companies such as Samsung and LG[4] have been the frontrunners. Car manufacturers such as Audi have incorporated OLED into their cars, the tail lights in their latest A8 release (July 2017) being made of OLED[5]. Google also has a foot in organic electronics, specifically targeting e-textile[6]. However, the research field with the currently biggest attention lies in medical patches used for health/fitness monitoring. L'Oréal[7] and Swedish Acreo[8] are two of the many companies developing/researching in this area.

CHALLENGES

While some fields within organic electronics have reached the stage where they can be mass produced, such as printed electronics, others remain at the infancy stage.

One of the benefits of organic devices is versatility providing endless opportunities. In general, the organic semiconductor can be manipulated such that different chemical and physical properties are obtained. This can be achieved by either manipulating the physical morphology, such as thickness, surface topology and similar, or by modifying the molecular structure of the materials through synthesis. Such a process is often too costly and too complex to be a relevant way to obtain the desired properties. How to construct and utilise different polymers and other organic materials to achieve different properties is a challenge which attracts a great attention from the research community.

Another challenge limiting the development is longevity, which we see in OLED technology. Compared to non-organic LED, OLED tends to be short-lived and may therefore need replacement more often; similar behaviour

is also seen in other organic electronics. However, production of organic devices is much cheaper than non-organic ones, which reduces the magnitude of this problem.

Furthermore, one of the major obstacles in the area of organic electronics is that stability of the devices is not always guaranteed. As an example, some organic devices are not stable in aqueous solutions which makes them not suitable for internal medical applications. This however, can be modified by changing the organic semi-conductor used in the device. On the other hand, this change may introduce new properties that are not desirable. Tuning and finding such properties is today one of the hottest research topics.

Due to these challenges, the field of organic electronics is far from having reached its full potential. However, as we could see in the case of OLED, when the technology is mature enough, many companies will be ready to get on the train. This shows huge potential in terms of interest and possibilities, particularly the unique properties such as flexibility, which would be desirable in many products. The manufacturing methodology that would become possible with printed electronics would give production of electronics the biggest change from the methods that we have today.

As previously mentioned, the driving force for the research in the area is the introduction of the devices in different environments and applications. For the devices to be commercially useful, key factors are stability, versatility, availability, longevity and production feasibility. Developing materials and devices that fulfil these requirements involves a deep knowledge in material science and chemical structures and their properties.

How to incorporate the same properties that we have today in conductive (non-organic) materials, into organic electronics, has yet to be solved for the majority of the cases and is one of the challenging areas that needs to be addressed.

TYPICAL INHERENT TECHNOLOGIES

With organic electronics embedded in all kinds of applications, we are looking at the need of massive-scale device-lifecycle management. This is also in line with the IoT vision that many companies have regarding the forecast of an explosive increase of IoT devices. Such device-lifecycle management would include management of industrial applications, just as well as medical applications, which in turn enables a massive scale autonomous environment.

Medical applications

Wearables in medicine (health monitoring)

This is one of the major applications for organic devices. The wearables within medicine often require the material to be soft, non-invasive and cheap to produce. These characteristics reinforce the requirements on availability and production feasibility. Example of such wearables include:

- **Oximeters:** For medical purposes, organic material can be incorporated into wearables such as pulse oximeters. The conventional oximeter is used for measuring a patient's pulse rate and arterial blood oxygen saturation with optoelectronic sensors composed of two light-emitting diodes (LEDs) with different peak emission wavelengths. The non-organic pulse oximeters are both expensive and rigid. The rigidity restricts the sensing since it cannot fully cover the finger on which it is placed. Organic electronic materials such as OLED with OPDs (organic photodiodes) allow softer oximeters to be used; the material is then bendable making the device cover the entire finger and hence improve the measurements while making it more comfortable for the patient. The organic LEDs and photodiodes are much cheaper, making the organic pulse oximeter far less expensive than conventional ones, hence demonstrating the possibility of cost reduction within health care.
- **Materials used to measure body fluids** Such as: sweat chloride to diagnose cystic fibrosis (CF); urine analysis to screen for help diagnose and/or monitor several diseases and conditions, such as kidney disorders or urinary-tract infections (UTIs); synovial sample from between joints to help diagnose the cause of joint inflammation, pain, and/or swelling.

Embedded drug delivery

In certain medical conditions, such as epilepsy or other disorders where the neural paths are perturbed, organic electronics have proven a useful tool for therapy and drug delivery. These devices are often in form of implants which require surgery; hence they put a high demand on stability (specially in aqueous solutions) and longevity. Some examples of such systems are:

- **Pain therapy:** Gamma-aminobutyric acid (GABA) is the chief inhibitory neurotransmitter in the mammalian central nervous system. It plays the principal role in reducing neuronal excitability throughout the nervous system. In humans, GABA is also directly responsible for the regulation of muscle tone. In cases where chronical neuropathic pain has been shown to be related to damaged nerves and hypersensitivity, there is evidence that the signalling has become perturbed and that a dysfunctional GABAeric system is related to the occurrence of spontaneous pain. Hence, by adjusting the level of GABA the pain may be controllable.
- **Epilepsy treatment:** There have been in-vivo studies on rats, where several models of the epileptic activity have been used to evoke epilepsy in the rat's brain. Epileptiform activity can be evoked by using pharmacological manipulations. These induced brain acitivites are the same type of activities that occur in the brain of a drug-resistant patient during an epileptic seizure. An organic electronic ionic pump (OEIP) has been used to pump ions into the desired area in the brain of the rat. The brain activity is measured by using a tungsten electrode and determining the cell-firing within the hippocampal slices. A high amount of cell-firing indicates high activity which is related to epileptic activity. By delivering GABA into the cells, it is shown that the cell-firing can be significantly reduced after one minute of pumping in ions and hence, the epileptic activity is reduced.

Industrial applications

Intelligent materials

- **E-skin:** Organic electronics can also provide augmentation of the human skin, for instance patches that can measure external parameters such as UV rays to estimate skin damage. Such patches can be created using organic electronic material. The patches can then collect the data and inform the user if they should get less exposure to the sun. L'Oréal has been developing such a patch with an associated app. They can also measure user parameters such as heart rate or oxygen concentration in blood, for instance during surgery. One advantage of using e-skin in these applications is the possibility of visualisation on the "skin", allowing for both data measurements and displaying of the data directly on the skin without relying on a mobile screen.

- **Smart textiles (e-textiles):** Smart textiles are textiles with incorporated electronics in the fibres, such as woven sensors or actuators. Another form of e-textile is a textile woven with conductive fabrics. The material must be drapeable (for instance when used on furniture) and conformable (for wearing).

For e-skin and e-textile to be commercially available, production feasibility and comfort are key factors.

Printed electronics

By turning an organic material, for instance a semiconductor, into "ink" or a "paste", it can be used to print circuits. This technique can be used to reduce manufacturing costs of circuits and also opens for home usage. The electronics can be directly printed on paper, and can be incorporated into the paper. When combined with additive manufacturing, the technique allows for electronic devices to be printed as a whole.

Electronic plants (e-plants)

The hormones, nutrition and other stimuli are transported through the plant xylems and phloem vascular circuits. The signals trigger, modulate and power different processes in the organism. By the controlling and measuring of these signals, changes in the physiological state of the plants are possible. This can be performed by the introduction of conductive material in the structure of the plant.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

Both medical and industrial applications require significant studies on possible body reactions to such devices before enablement, specifically the applications that are integrated into, or placed on, the body. For instance, electronic skin placed on top of the skin would prohibit the skin from absorbing air which in turn would cause inflammations when worn for longer periods of time. Making the various devices survive long-term is a key factor for organic electronics. The same importance applies to the self-repairing of organic devices; optimal performance would be when the material and the body (host) do not only co-exist, but collaborate. An example would be if the body could repair damage to the integrated e-skin.

Another breakthrough that would result in enablement for organic devices is continuous charging. An optimal method would be where the device efficiently utilises its surroundings to power itself. In a device that is body-worn or integrated into the body, the ideal case would be to use natural body movement to power itself, such as the beating of the heart, or the chest expansion when inhaling to generate power.

With more and more devices being created, an autonomous way of communication and visualization is needed. With the same reasoning, aside from connectivity, a system that enables massive-scale device-lifecycle management is necessary. Breakthroughs in IoT that specifically target such matters would be essential for organic electronics as well.

Potential for Swedish positioning

Sweden's position at the forefront in both medical and industrial research creates a solid foot-hold for new technology, such as organic electronics. This comes from the Swedish welfare model that allows substantial risk-taking for innovation. As a consequence, various tech start-ups in Sweden are rising up, which shows an interest in exploring new technical solutions. Similarly, large players with both medical and industrial background would most likely approve of this topic as vital for the digital transformation. These factors indicate that Sweden could be a profound base for new technology in this area.

POTENTIAL PLAYERS

In general:

- Google will most probably start developing and researching the topic as they are involved in many medical patents and research. They also focus on patents related to body augmentation such as lenses for screen projection and temperature measuring patches.
- Microsoft will probably be involved in organic electronics as part of their medical and wearable research and development.
- Many of the major medical companies will start using the research conducted in areas of drug delivery and measurement and sensors to provide new solutions and products for hospitals and major customers. The probability of them already conducting research is low, but there is a possibility of them co-working with universities and research institutes.
- We are already seeing the rise of many start-ups and small companies creating prototypes of wearables and devices for customers/individuals. These start-ups may be acquired by "giants" such as Google/Microsoft and similar.

Electronic plants

Linköping university in Sweden has researched and developed methods for stimulating and regulating plants physiology[9]. With a Swedish contributor already working in this research area, many other partners can collaborate to establish solid ground within the field.

Wearables in medicine

Sweden has many players that could contribute to this area. Combining the knowledge of the frontrunners within each field would allow wearables in medicine to be developed on multiple fronts simultaneously, targeting one common goal. The RISE research institute in Sweden is heavily involved in organic electronics.

Organic electronics in IoT environments:

- Ericsson is working with many partners and has a profound Swedish background in terms of connectivity for the sake of autonomous communication and visualization; they therefore have good prerequisites to play a role that will allow the technology to develop fast.
- Organic electronics can be used to verify food quality, as well as medicine supplies quality, during both container shipment and storage. In Sweden, Astra Zeneca, Scania, Volvo and Karolinska institute, are just a few of the contributors that could serve a vital role in such research and development.

NEED FOR SUPPORTING ACTIVITIES

Current players in Sweden, companies, research institutes and universities, have all excelled in their different fields. However, for organic electronics to rise, the research has to be tied together somewhere and that is what is lacking. There are no bridges between the fields which would give rise to this new direction.

Collaborations between such Swedish players, universities and organizations that have established themselves, would allow organic electronics to be built around an autonomous system of future devices.

Footnotes and links:

1. Coatanéa, E., Kantola, V., Kulovesi, J., Lahti, L., Lin, R., & Zavodchikova, M. (2009). Printed Electronics, Now and Future. In Neuvo, Y., & Ylönen, S. (eds.), Bit Bang – Rays to the Future. Helsinki University of Technology (TKK), MIDE, Helsinki University Print, Helsinki, Finland, 83-102.
2. Printed Organic and Molecular Electronics, edited by D. Gamota, P. Brazis, K. Kalyanasundaram, and J. Zhang (Kluwer Academic Publishers: New York, 2004).
3. www.printedelectronicsworld.com/articles/7830/printed-temperature-sensor-for-drug-monitoring-in-field
4. www.lgdisplay.com/eng/prcenter/newsView?articleMgtNo=4962
5. www.oled-info.com/new-audi-a8-sports-oled-rearlights
6. Pacelli, M., et al. “Sensing fabrics for monitoring physiological and biomechanical variables: E-textile solutions.” Medical Devices and Biosensors, 2006. 3rd IEEE/EMBS International Summer School on. IEEE, 2006.
7. www.bbc.com/news/technology-35238636
8. www.acreo.se/expertise/printed-and-organic-electronics
9. www.liu.se/en/article/liu-researchers-create-electronic-plants

Radio technology

Hugo Tullberg, Ericsson Research, hugo.tullberg@ericsson.com

OVERVIEW

This concerns research on autonomy and machine learning in the context of global information and communication systems. Such systems are becoming increasingly complex and heterogeneous, and the requirements put on them are becoming increasingly diverse. Manual deployment, configuration, optimization and maintenance will not be possible in the near future, and increased levels of autonomy are required to meet all demands. The ultimate benefit of autonomy is increased service levels, quality, and reliability at decreased effort and cost.

INTERNATIONAL MATURITY

The global ICT (information and communication technology) networks per se are well established. However, ICT systems are continuously applied to new areas in the ongoing digitalization of industry and society, and the technology development is extremely rapid. Though there exist pockets where autonomy and machine intelligence have reached far, the general level of autonomy is still low.

Leading players in the cellular communication domain include Ericsson and other vendors. Terminal and device manufacturers include for instance Sony, Apple, and Intel. Service providers include for instance Google, Apple, and Facebook. Early adopters of Industrial IoT (Internet of things) include for instance ABB, Scania, Volvo, and Siemens. Research in ICT-related autonomy and learning is not as visible as research on vision, natural language processing, and similar, though research groups are starting to emerge.

CHALLENGES

The main challenge is the rapid diversification of previously comparably homogeneous systems. This change is resulting partly from a merge of what was previously separate systems and partly from introduction into new application areas. The challenge manifests itself in:

Diversity in requirements

Emanating from for instance:
- AR/VR (augmented reality/virtual reality) – typically requiring very high data rates, low latencies to avoid dizziness, and high computational capabilities (likely requiring edge cloud);
- evolution of MBB (mobile broadband) – typically requiring high data rates but not as low latencies as AR/VR;
- massive machine-type communication (mMTC) addressing low-capacity devices as sensors – typically requiring very low power consumption but tolerating high latency and low data rates;
- ultra-reliable and low-latency communication (URLLC) – (requirements are self-explanatory); and
- whatever the future brings.

Diversity in communication modes

For instance:
- point-to-point communication – to connect two points to each other;
- cloud solutions for “always same experience” – to provide a consistent experience while on the move;
- point-to-multipoint communication – to broadcast, for instance a sporting event.

Diversity in radio interfaces

Different radio interfaces have different characteristics which are more or less well suited to support different applications, and also have different network architectures. For instance, GSM is heavily used for MTC traffic today and has good coverage, whereas NR will support high data-rates and low latency but will have a coverage challenge at higher frequencies.

Diversity in capabilities

Depending on storage and compute capabilities, we can apply different types of machine intelligence at different network nodes, typically:
- data centers;
- network nodes;
- terminals;
- devices, sensors and wearables.

Together with new trends in network architecture, this results in a network where communication, storage and computation capabilities are spread and dynamically assigned to users and services. Within a near future it will be impossible to deploy, configure, optimize and maintain without autonomy and intelligence.

TYPICAL INHERENT TECHNOLOGIES

The research topics manifest themselves in several areas, including:

Service domain

- Service flows for point to point communication.
- Adaptive transfer of storage and compute resources between network nodes for ASE (area-spectral efficiency).
- Response-time optimization.

Network domain

- Communication selection, cognitive network.
- Coverage and cell optimization.

Communication domain

- Intra-RAT (radio access technology) optimizations.
- Lower protocol layer optimizations.
- Trade-off between communication data processing and learning data processing.

This document concerns research topics into the communication domain with a particular focus on physical layer and radio-related topics.

POSSIBLE ENABLING RESEARCH BREAKTHROUGHS

In the network domain, topics related to load balancing and traffic optimization can significantly improve end-user experience, communication reliability, and network utilization. Autonomous cell shaping and beam selection can improve throughput. Understanding of traffic and mobility patterns can both improve network utilization and quality of experience, as well as detect undesirable usage. The overall goal is to make use of spatio-temporal properties on all scales to transform the network and traffic management from reactive to predictive.

In the communication domain, the radio-signal-processing algorithms are often perceived as close to optimal. However, the optimality hinges on specific models and assumptions, of which some are made to make real-world problems analytically tractable. As computation power increases, we can take a more data-driven approach, and potentially apply machine learning to a range of problems. It should be investigated which assumptions can be removed and models be replaced by real data or made significantly more realistic.

Breakthroughs in these areas can lead to a better utilization of scarce radio resources, and an ability to better meet diverse needs.

POTENTIAL FOR SWEDISH POSITIONING

Sweden has a long tradition of world-leading research and innovation in wireless communication, industrialization and manufacturing. Sweden has a high ICT maturity and well developed infrastructure. Sweden thus has the necessary actors and competences to pose relevant research questions for machine learning in communication systems. This allows Sweden to quickly utilize and commercialize research results once relevant machine learning results are available.

POTENTIAL PLAYERS

Communication research: Ericsson, Saab, SICS, and relevant universities.

Industrial IoT application: ABB and Volvo.

NEED FOR SUPPORTING ACTIVITIES

The overarching question is what gains can be expected when replacing today's highly optimized algorithms with data-driven autonomous solutions. Learning theory is a well-established field; however, theoretical bounds on achievable performance of machine learning (a learning-theoretical counterpart of Shannon theory) should be established to assess potential gains when assumptions and models are replaced by data-driven approaches. Theoretical assessments of gains under practical assumptions, for instance limits on available time, computational power, see below, should also be made.

Specific algorithms, for instance beam selection, beam design, any L1 (layer 1) algorithm where data is abundant and assumptions are made (for example on fading characteristics, noise densities) is another area where activities would be welcome.

There is also a need for undertakings within distributed learning and reduced resource learning. Deep learning in centralized data centers depends on great computational power, availability of very large data sets, sufficient time for learning, and to some extent on high-precision arithmetic. Challenges to move this into the communication network include reduction of the aforementioned factors. Research topics include pre-learning before deployment, real-time learning ("on-the-job training"), learning from partial data sets, exchange of learning (without having to transfer gigabytes of learnt weights), and machine learning on limited-capabilities platforms.

Extensions to network learning targeting limited-autonomy units form a fully autonomous system, and autonomous units connected to autonomous systems. What should a limited-capability node learn and what should just be pushed onwards?

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

The Wallenberg AI, Autonomous Systems and Software Program (WASP) is a major national initiative for strategically motivated basic research, education and faculty recruitment in autonomous systems and software development. The ambition is to advance Sweden into an internationally recognized and leading position in these two areas.

The starting point for WASP is the combined existing world-leading competence in Electrical Engineering, Computer Engineering, and Computer Science at Sweden's four major ICT universities: Chalmers University of Technology, KTH Royal Institute of Technology, Linköping University, and Lund University. WASP will strengthen, expand, and renew the national competence through new strategic recruitments, a challenging research program, a national graduate school, and collaboration with industry.

The scope of WASP is collaborating vehicles, robots and complex software-intensive systems with the intelligence to achieve autonomy in interactions with humans. Software, models and algorithms are currently a large and rapidly increasing part of the development of almost all engineering systems including autonomous systems, and there is a strong need to manage this complexity to ensure functionality and reduce development costs. Autonomous systems are a scientifically challenging disruptive technology that will fundamentally change society and industry. Swedish industry is by tradition strong in systems engineering. To stay competitive, Sweden needs to invest in research and competence building in this area, and hence WASP is highly relevant to Swedish industry.