EIT Digital – Industrial PhD position proposal

PhD thesis information

PhD Thesis – Title		Network technologies for big data applications
PhD Thesis – Short summary	Max 100 words	The massive deployment of IoT, the embedding of sensors in any object, the rise of smart materials in the next decade are all sources of data representing a growth of two orders of magnitude over the next 5 to 7 years. In order to deploy network services collecting and processing data efficiently, we need to use virtualized software on top of hybrid platforms (e.g. Edge/Fog/Cloud computing) serving both storage, processing and visualization (semantics extraction). The aim of the thesis is to model, analyze and design high performing big data frameworks taking advantage of distributed (in geographical sense) systems and are reliable for applications that will characterise the everyday life of society in the near future.
Rationale/challenge – *describe the problem and why it is relevant*	Max 200 words	Today big data analytics software can handle live data streams quickly and efficiently, the latter in terms of available computing resource utilization and process scheduling. However, these solutions typically do not consider the performance of the underlying network, that may become a bottleneck. With the help of SDN and NFV, data analytics functions can be virtualized (e.g. virtual machines and/or containers). This leads to extra computing overhead but offers the possibility to tackle the performance issues of the underlying networks. Furthermore, with such technologies, we have the opportunity to place virtualized big data VNFs as close as possible to the data sources decreasing the bandwidth load of the network. This approach can provide higher reliability to the deployed service by monitoring, scaling, healing and migrating VNFs in the virtualized platform. In order to apply these functionalities we need to use stateless VNFs, which means that VNFs and industrial IoT systems can externalize their states to multiple low-latency shared memory systems. There are numerous questions related to the control and data planes of the big data analytics VNFs that have to be solved in order to apply such approaches and deploying analytics services optimally in a geographically scattered infrastructure.
Innovation – *describe what is the intended solution and the advance w.r.t. the state-of-the-art*	Max 250 words	The huge amount of data generated that can be harvested through 4G and 5G networks and to the spread of public computing infrastructure, such as cloud and fog platforms, may support the rise of new applications. Some of these may require very challenging bandwidth and latency parameters that can only be addressed through an orchestration of network resources from the edges and the implementation of flexible architectures. The goal of the thesis is to provide placement solutions for data stream analytics applications, which take into account the network characteristics during the implementation of virtualized data analytics services and also enable high reliability using healing, scaling and other supporting capabilities. The proposed work includes: · defining use-cases for virtualized live data analytics

		services in the fields of traditional Telco and 5G services · determining the characteristic differences of current virtualization platforms (OpenStack, Docker, etc.) for big data services in geographically distributed scenarios · clarifying the disadvantages of current big data applications (e.g. Spark, Storm, in-memory Key-Value stores) since they don't take into account network limitations · examining existing big data resource managers (e.g., YARN, Swarm, Kubernetes) and designing appropriate scheduling and placement solutions · evaluating different data plane solutions (SDN switches, DPDK) to provide efficient networking for big data VNF components · evaluating different database solutions which could provide stateless approach for the virtualized software functions and define a n optimal placement policy for them · implementing an integrated solution that takes into account the underlying network and deploys data analytics services optimally in terms of performance and resource usage During the research, the use cases will be decomposed to process and storage components, and the operations will be investigated from both control and data planes perspectives. Evaluation and validation of the design will be tested on bare metal and on virtual resources as well.
Research focus/topics – *describe <u>how</u> you are going to solve the problem*	Max 200 words	The investigation areas can be grouped into four phases that must be completed during the PhD program. - The first phase is the clarification of current and future use cases: determining their requirements and creating an overview on the state-of-the-art of theoretical and implemented solutions to high-performing big data stream analytics applications. To this end, a unified framework for requirements is to be designed and a comprehensive study of the capabilities of related work is necessary. - Second, the applicant models analyzes, designs, evaluates, and implements prototypes of the proposed new concepts. This task forms the bulk of the project work, and the applied methods include mathematical modeling of distributed systems, practical methods of code development and deployment. - In the third phase, the proposed systems are studied through comprehensive simulations and measurements in real-life deployments; tests will be performed, both in small-scale and large-scale environments. The creation of a benchmarking framework for the evaluation is mandatory, and this task runs iteratively with the design and implementation task. - Finally, leveraging on the results, the proposed solutions are to be further improved for a few well-specified use

		cases to reach the highest achievable performance and efficiency in the process of the data analytics service deployments in those areas. This task creates a strong link to the industry, and ensures the practical applicability of the results of this research project. In order to have the opportunity to deploy live big data services on a large-scale, a strong cooperation with Ericsson is ensure through the whole activity.
Deadlines/milestones (Gantt chart)	M6	Review the state-of-the art, surveying best practices, frameworks, guidelines and determine the performance requirements of big data stream analytics applications. Take notice of the industrial and real networks constraints. Find industry use cases that provide basis for delay sensitive live-stream applications scenarios. Publish the findings.
	M12	Implement first prototype of the identified solutions. Analyze it by simulations and small-scale tests. Publish the findings. Determine a possible solution to provide stateless working for VNFs with low-latency in line with real network constraints.
	M24	Identify a solution for stateless working of VNFs to provide monitoring, debugging and alerting options for deployed services. Publish the findings.
	M36	Extend and deploy the prototype to support multiple virtualization and network platforms on which it is orchestrated. Test different compilations and determine the most effective one. Transfer technology to Ericsson. Publish the findings.
	M48	Optimize the prototypes based on the previously gathered performance experience and on real world verification in cooperation with Ericsson. Compare the results with other available solutions. Summarize the results and complete the PhD thesis.
Expected outcome – *describe the expected results of the PhD*	Max 100 words	The expected results of the PhD are solutions to enable new big data analytics placement algorithms that provide live analysis, the possibility of function migration, optimal orchestration of services and reliable working process of the deployed services. The results shall include: - Comprehensive study of the role of the networks in terms of orchestrating and deploying big data analytics system components - Detailed analysis of the impact of network performance on data analytics systems - Prototype of big data stream analytics system for a virtual platform deployed on a distributed infrastructure tested in real life environment in cooperation with Ericsson. - Published papers in high-quality academic conferences and journals - Implementations that are usable in core and 5G Telco products of Ericsson

Relevance for the Action Line (section to be filled out by the Action Line Leader)

Action Line	AL	
Alignment with Action Line – *statement from the Action Line Leader indicating the relevance for the AL from his perspective*	Max 150 words	This research topic is definitely in line with the focus area of the Action Line. Effective distribution of processing is important for several reasons like the latency in processing, cost of transmitting data but also energy consumption that could be taken into account when optimizing and orchestrating the actual processing with the network characteristics.
Relevant IA – *List any relevant Innovation Activity (if applicable)*	Max 100 words	During 2017 we had a project called Hopsworks that was focusing on providing a platform for near-stream analytics. It would be very nice if the PhD student would participate in potential future IAs with investigations, workshops and similar.

Partnership/financial information

Action Line Leader	Name	Henrik Abramowicz
Industrial partner		Ericsson
Industry advisor – *name and short bio*	Max 100 words	Péter Mátray received his M.Sc. (2005) and Ph.D. (2014) in computer science from Eötvös Loránd University (ELTE). During his work at the university he was focusing on topics of large-scale active Internet measurements (e.g., IP-geolocation), and the challenges related to the efficient management and analysis of massive measurement data sets. He joined Ericsson Research in 2012, where he has been involved in various projects to create systems such as, an analytics platform for customer experience management, a data-driven monitoring framework for troubleshooting complex cloud applications, and recently, a low-latency data sharing system for 5G and industrial clouds. Currently Peter does research in Distributed Computing and Computer Communications (Networks). Their most recent publication is 'DAL: A Locality-Optimizing Distributed Shared Memory System.
Academic/research partner		Budapest University of Technology and Economics (BME)
Academic/research supervisor – *name and short bio*	Max 100 words	Laszlo Toka is currently working as an Assistant Professor in the High-Speed Networks Laboratory at the Department of Telecommunications and Media Informatics at the Budapest University of Technology and Economics, Hungary, where he received the M.Sc. and Ph.D. degree with summa cum laude in computer science in 2007 and 2011, respectively. Between 2011 and 2014 he worked as research fellow at Ericsson Research. In 2014 he won the Janos Bolyai Research Scholarship of the Hungarian Academy of Sciences. His research interests include big data, network economics, SDN and cloud.
HEI granting the title		Budapest University of Technology and Economics (BME)
DTC location	Node	Budapest
Geographical mobility plan		
No. of PhD positions	[#]	1

PhD duration	[#years]	4 years (the official duration of the Hungarian PhD state scholarship)
Co-funding percentages: - Industry	[%]	20%
- Academia	[%]	30%
- EIT Digital	[%]	50%