

# HelmGemm: Managing GPUs and FPGAs for transprecision GEMM workloads in containerized environments



## HelmGEMM

Dionysios Diamantopoulos, Christoph Hagleitner  
Heterogeneous Cognitive Computing Systems Group  
IBM Research – Zurich  
(did,hle)@zurich.ibm.com

July 15-17, 2019, Cornell Tech, New York

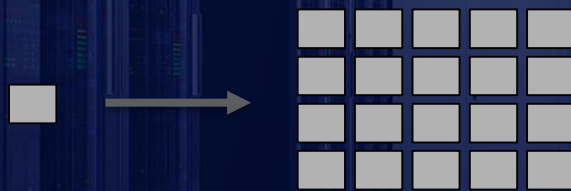


# Homogenous was yesterday's approach

## *The AI era requires a new one*

### Legacy Approach

ONE SIZE FITS ALL -  
Approach all application requirements with a single non-optimized building block




### Modern Approach

*Leverage optimized servers designed for the AI era and the vastly different requirements*




# Typical use-cases in need of advanced computing

**Automotive, Transportation and Logistics**




- Autonomous driving
- Pedestrian detection
- Accident avoidance
- Predictive Maintenance
- Digital twin
- Logistics optimization

**Security, Public Safety and Traffic control**




- Video Surveillance
- Image analysis
- Facial recognition
- Predictive crime
- Traffic prediction
- Cyber Security

**Consumer, Web, Mobile & Retail**




- Image tagging
- Speech recognition
- Natural language
- Sentiment analysis
- Recommendation
- Social analysis & trends

**Broadcast, Media and Entertainment**




- Captioning
- Search
- Recommendations
- Real time translation
- Consumer behaviour

**Medicine and Biology**

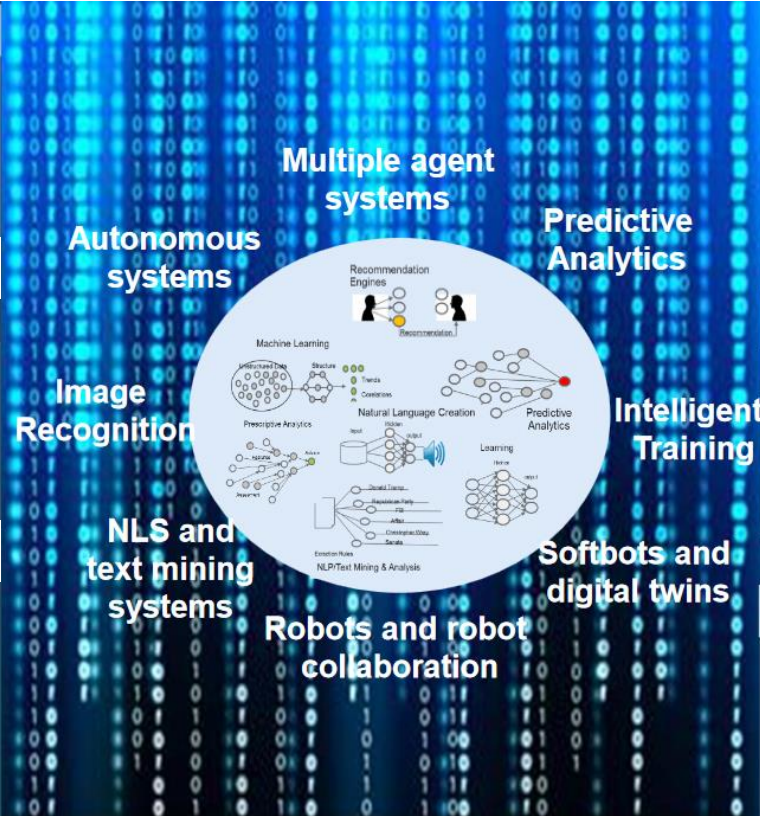


- Drug discovery
- Diagnostic assistance
- Cancer cell detection
- Brain research
- Genome research
- Field studies

**Banking, Finance & Insurance**



- Trend prediction
- Document analytics
- Recommendation
- Service & Chatbots
- Trading forecast
- Risk management



**Multiple agent systems**

**Autonomous systems**

**Predictive Analytics**

**Image Recognition**

**NLS and text mining systems**

**Robots and robot collaboration**

**Softbots and digital twins**

**Intelligent Training**



# Systems designed to crush Big Data and AI workloads

- Deep Learning



- Data Intensive Workloads



- Big Data Workloads



- Enterprise Private Clouds



# Acceleration options for POWER™

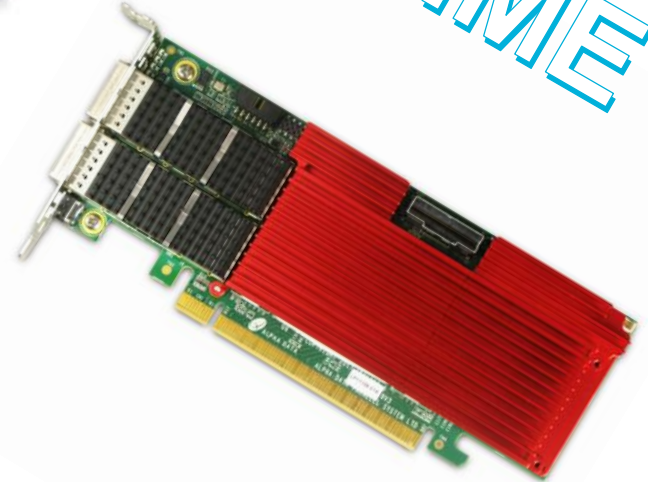
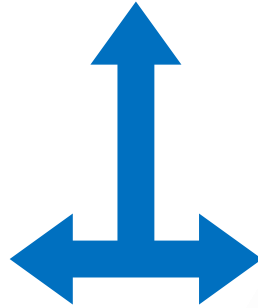
FAST

GPU



FPGA

REAL  
TIME



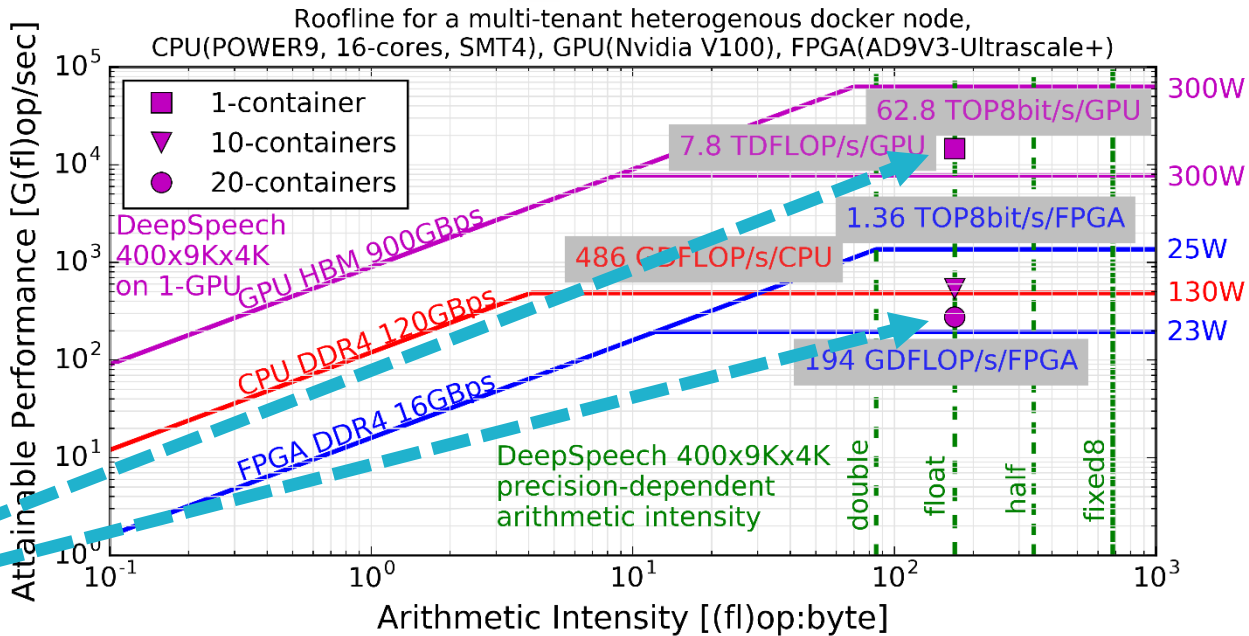
- Thousands of tiny CPUs using high parallelization
  - compute intensive application
  - SIMD-oriented workloads

- Logic + IOs are **customized** exactly for the application's needs.
- Very low and **predictable** latency applications
- MIMD-oriented workloads



# HelmGemm Motivation

*The observation!*



**-53x**

GPU memory sharing in containerized systems can lead to GPU performance inefficiencies that fall within the performance envelope of FPGAs, which operate on a power budget one order of magnitude lower.

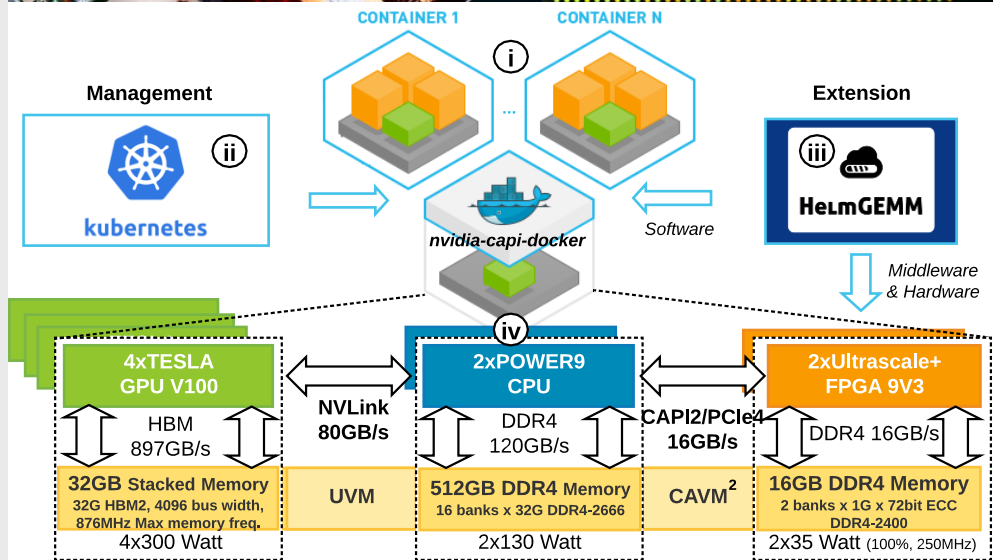
# HelmGemm components

**Docker container service:** multi-tenant environment with a high-level API to provide lightweight containers that run processes in isolation

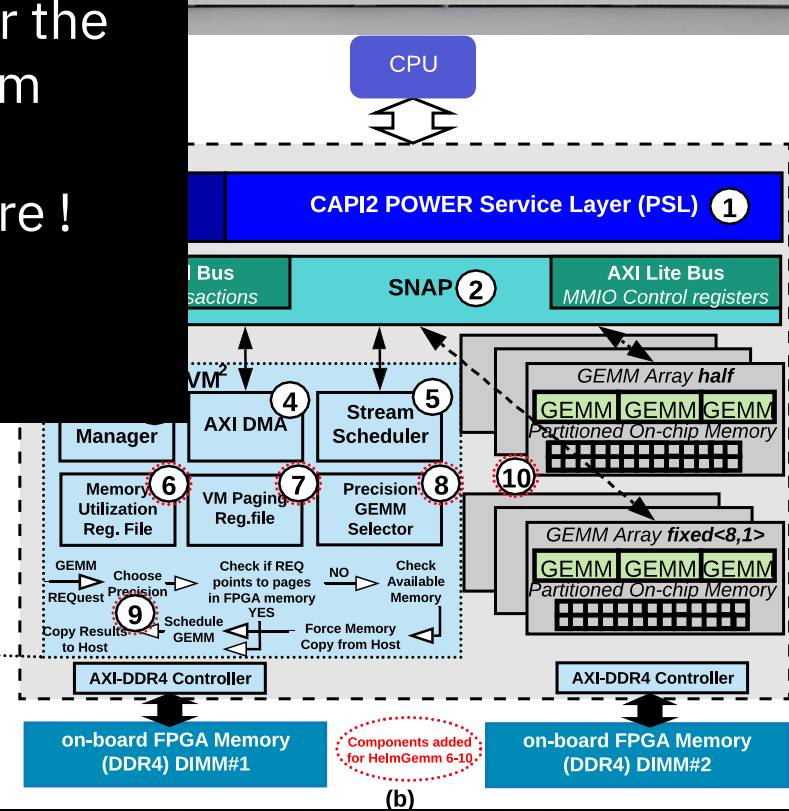
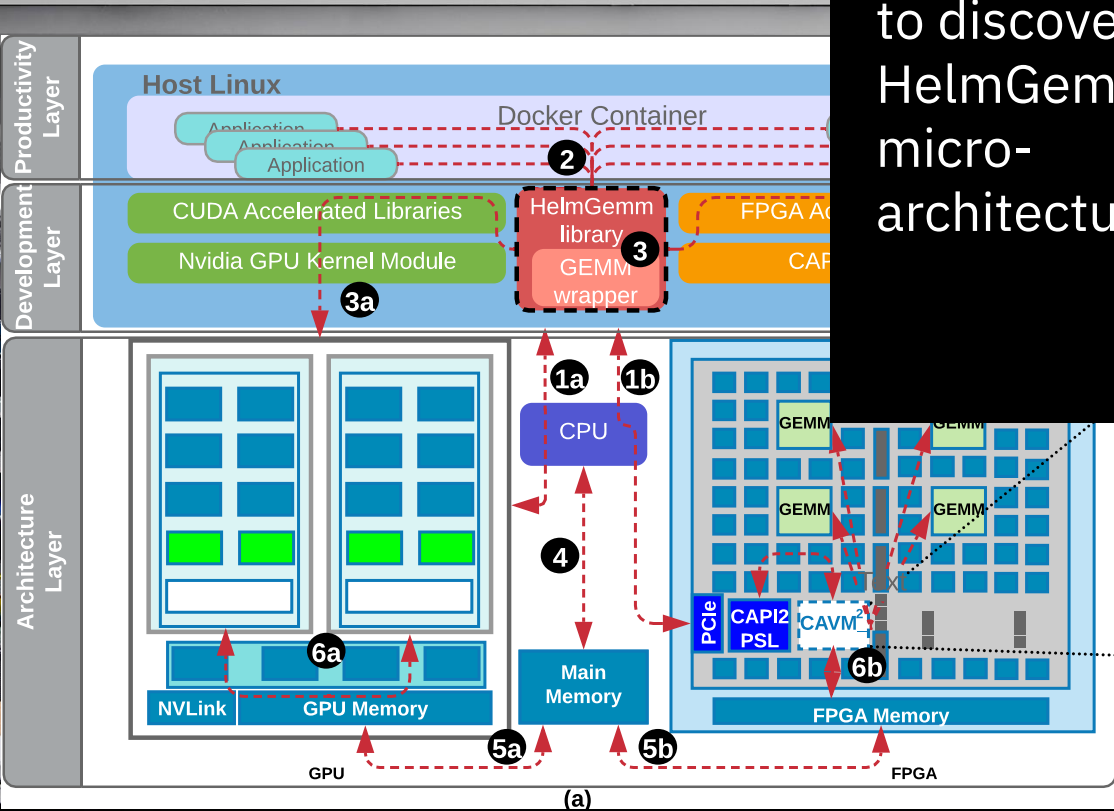
**Kubernetes management:** deploy, maintain, and scale applications

**HelmGemm extension:** hardware, middleware and software

**Hardware support :** 4xGPUs, 2xFPGAs



Visit our poster to discover the HelmGemm micro-architecture !





**POWER9 AC922**



**POWER9 AC922 + V100 + 9V3**



**59.3x** more performance



**28.7x** more energy efficiency

# Thank you

Dionysios Diamantopoulos  
Heterogeneous Computing Systems Group

—  
did@zurich.ibm.com  
+41-44-724-85-25




# oprecomp.eu

# Visit our poster

## HelmGemm: Managing GPUs and FPGAs for transprecision GEMM workloads in containerized environments

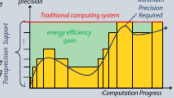
Dionysios Diamantopoulos and Christoph Hagleitner  
IBM Research – Zurich, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland  
(did, hlej@zurich.ibm.com)



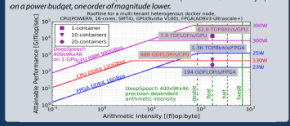
**Ambition** OPRECOMP project aims to build an innovative, reliable foundation for computing based on transprecision analytics (enabling the conservative 'precise' computing abstraction and replacing it with a more flexible and efficient one, namely transprecision computing. OPRECOMP milestones to achieve demonstrate this idea in the domains of:

- Big Data Analytics,
- Deep Learning,
- NPC Simulations

from the sub-bit to the Megabit range, spanning five orders of magnitude.

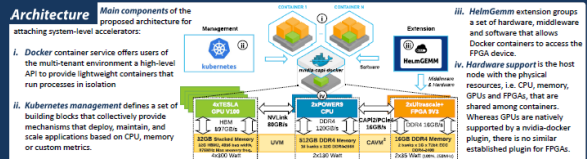


**HelmGemm Motivation** GPU memory sharing in containerized systems can lead to GPU performance inefficiencies that fall within the performance envelope of FPGAs, which may operate on a power budget, one order of magnitude lower.

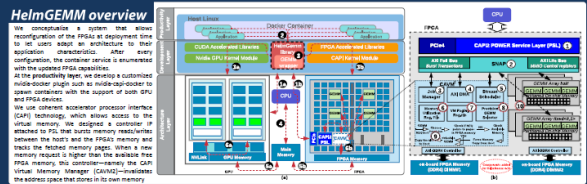


**Architecture** Main components of the proposed architecture for attaching system-level accelerators:

- Docker container service offers users of the multi-tenant environment a high-level API to provide lightweight containers that run processes in isolation.
- Kubernetes management defines a set of building blocks that collectively provide mechanisms that deploy, maintain, and scale applications based on CPU, memory or custom metrics.
- HelmGemm extension groups a set of hardware, middleware and software that allows Docker containers to access the FPGA device.
- Hardware support is the host node with the physical resources, i.e. CPU, memory, GPUs and FPGAs, that are shared among containers. Whereas GPUs are natively supported by a multi-docker plugin, there is no similar established plugin for FPGAs.

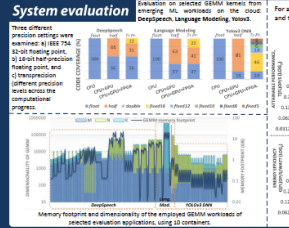


**HelmGemm overview** We conceptualize a system that allows reconfiguration of the FPGA or deployment time to set users about an architecture to their application. In practice, after every configuration, the container service is enumerated with the system FPGA capabilities. At the probability layer, we create a customized end-user high-level multi-tenant to spawn containers with the support of both CPU and FPGA services. We use coherent accelerator processor interface (CAPI) technology, which allows access to the virtual memory, we designed a controller in order to FC, we utilize memory resources between the host and the FPGA's memory and track the address memory map. When a user memory request is higher than the available free FPGA memory, the container-manages the Cache Virtual Memory Manager (CAVMM)—provides an address space that points to its own memory.




**System evaluation** Evaluation on diverse, domain specific, emerging ML workloads in the cloud and DeepSpeech, Language Modeling, Video CNN.

Three different precision settings were examined: IEEE754, 20-bit floating point, 10-bit and 1-bit transprecision. Meeting goals, and 1 transprecision operation offers a 10x performance gain across the computational program.



For all three applications, we have measured an average energy efficiency gain of 1.7x and 2.3x in half and transprecision settings, considering four settings as the baseline for every system setup.

In contrast to the lower performance of GPU/FPGA on design and given the lower power of FPGAs, the energy efficiency of the derived solutions increases at the system level.



This project is funded by the European Union's Horizon-2020-747177 (Project research and innovation program under grant agreement #72617). IBM and POWER are trademarks of International Business Machines Corporation in the United States, other countries, or both. Other product and service names might be trademarks of IBM or other companies.