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

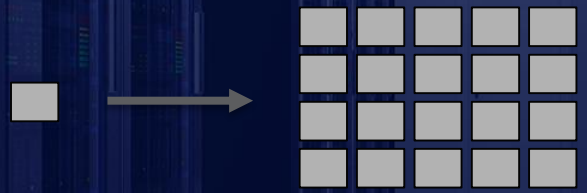
IBM

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

Multiple agent systems

Autonomous systems

Predictive Analytics

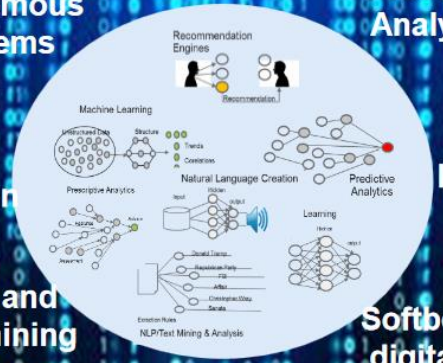


Image Recognition

Intelligent Training

NLS and text mining systems

Softbots and digital twins

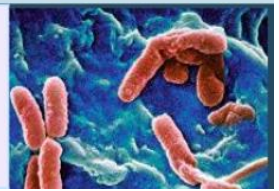
Robots and robot collaboration

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

Systems designed to crush Big Data and AI workloads

- Deep Learning

- Data Intensive Workloads

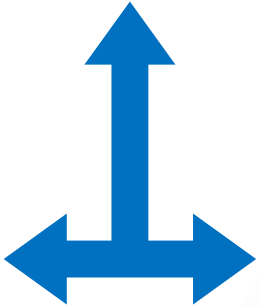
- Big Data Workloads

- Enterprise Private Clouds

Acceleration options for POWER™

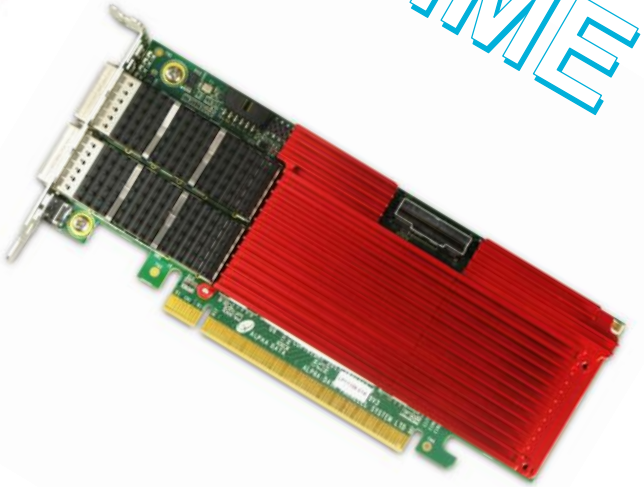
FAST

GPU



FPGA

REAL TIME

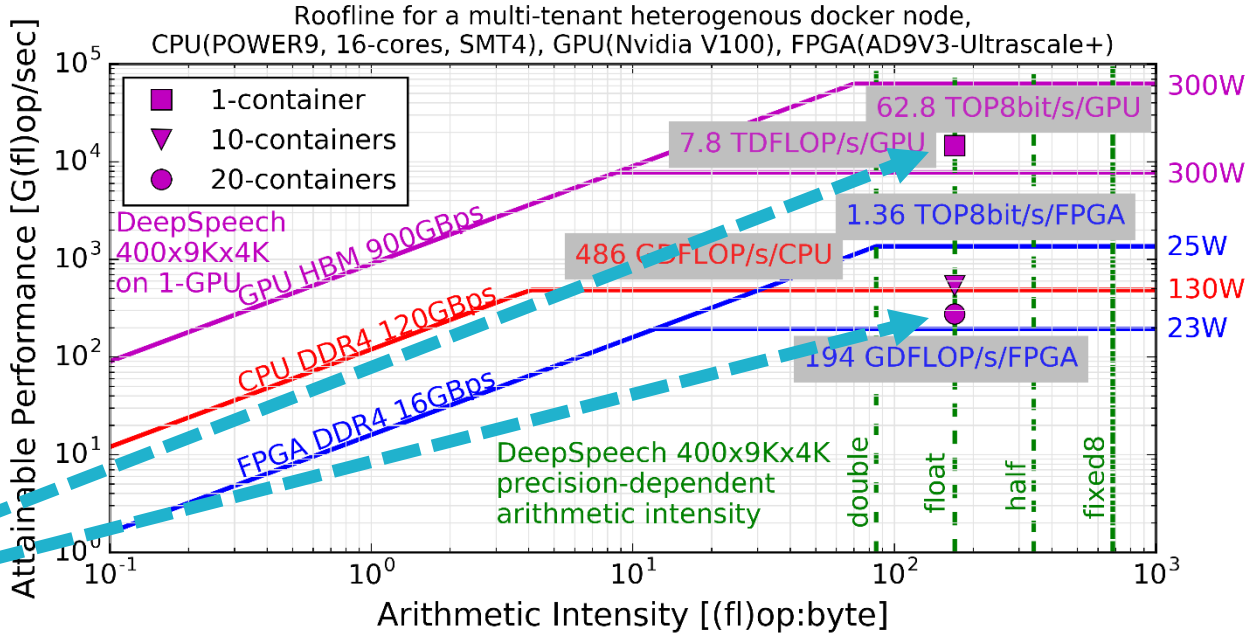


- Thousands of **tiny CPUs** using high parallelization
  - compute intensive application
  - SIMD-oriented workloads

- 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.

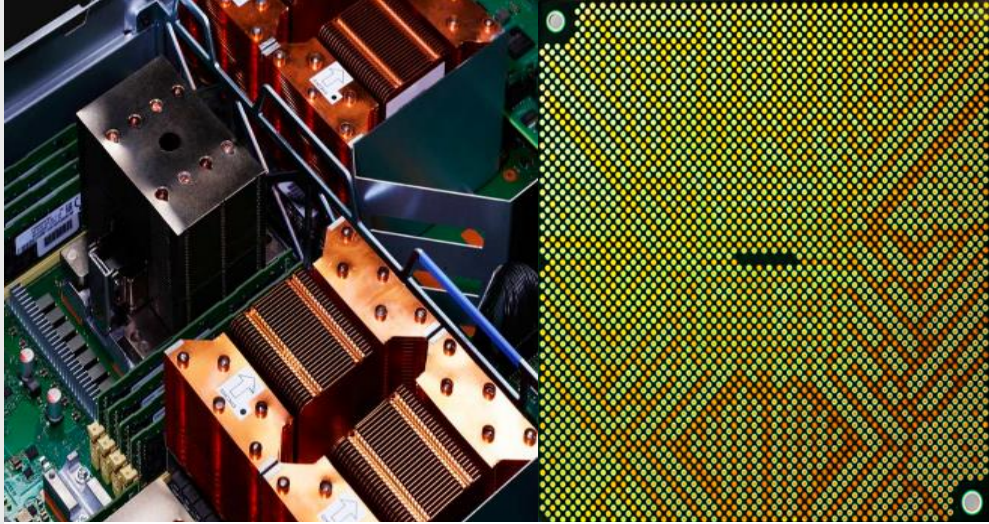
HelmGemm components

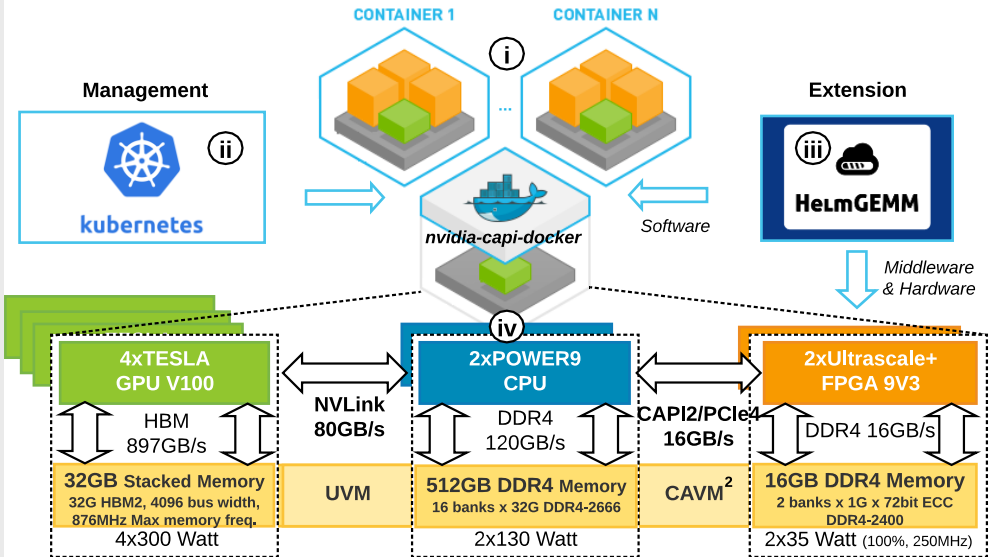
Docker container service: multi-tenant environment with a high-level API to provide lightweight containers that run processes in isolation

Kubernetes management: deploy, maintain, and scale applications

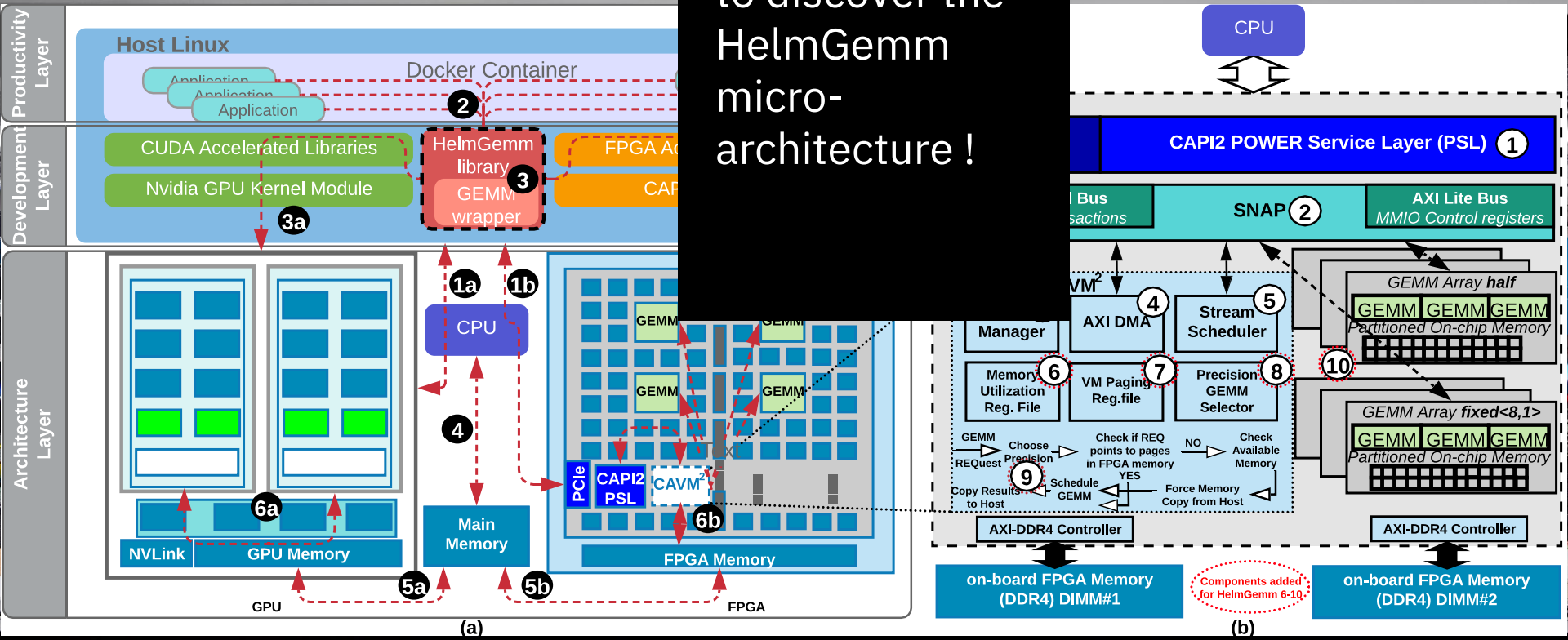
HelmGemm extension: hardware, middleware and software

Hardware support : 4xGPUs, 2xFPGAs





Visit our poster to discover the HelmGemm micro-architecture !



POWER9 AC922

HelmGEMM

POWER9 AC922 + V100 + 9V3

83 G(fl)Ops/sec

total operations per second

4.94 T(fl)Ops/sec

59.3x more performance

0.16 G(fl)Ops/sec/Watt

total operations per second per watt

4.78 G(fl)Ops/sec/Watt

28.7x more energy efficiency

Thank you

Dionysios Diamantopoulos
Heterogeneous Cognitive Computing Systems Group
—
did@zurich.ibm.com
+41-44-724-85-25

OPRECOMP
Open Transprecision Computing

oprecomp.eu

Visit our poster