Heterogeneous Computing Systems in Cloud Datacenters

Christoph Hagleitner, hle@zurich.ibm.com
IBM Research – Zurich Lab (ZRL)

- Established in 1956
- Two Nobel Prizes (1986 and 1987)
- Today
  - ~300 employees (~3000 worldwide)
  - 40+ different nationalities
  - open innovation w/ 277 projects & 1900 partners in FP7, H2020, ...
Acknowledgements

Accelerator Technologies @ ZRL

- μServer team @ ZRL (Martin Schmatz, Ronald Luijten, ...)
- Supervessel team
- openPOWER team
- Peter Hofstee, Alessandro Curioni, ...

L.Fiorin, F. Abel, E.Vermij, J.Weerasinghe, S.Dragone, M.Purandare, R.Polig, J.vanLunteren, H.Giefers, C. Hagleitner
IDC @ 2016: “Transform Or Die”

- IDC: 30% of today’s tech suppliers will not exist as we know them today, having been acquired or failed.
- IDC: 1/3 of the top 20 companies in every industry will be “disrupted” over the next 3 years, meaning their revenue, profits, and market position will deteriorate, not that they will go out of business.
- Forrester: “Lead The Customer Obsession Transformation”
- Forrester: Customers expect consistent and high-value in-person and digital experiences. The risks in today’s customer-led market have shifted from responding too early to responding too late.
- Gartner: “Autonomous software agents will play a crucial role in the economy and everyday life.”
- Gartner: The future will belong to the companies that can create the most effective autonomous and smart software solutions.
Where is the IT Industry going ...???

1. By the end of 2017, two-thirds of the CEOs of global 2000 enterprises will have digital transformation at the center of their corporate strategy.

2. By 2017, over 50% of organizations’ IT spending will be for third platform technologies (cloud, mobile, social business and big-data analytics), solutions, and services, rising to over 60% by 2020.

3. By 2018, at least half of IT spending will be cloud-based, reaching 60% of all IT infrastructure and 60-70% of all software, services, and technology spending by 2020.

4. By 2018, enterprises pursuing digital transformation strategies will more than double software development capabilities; two-thirds of their coders will focus on strategic digital transformation apps and services.

5. By 2018, enterprises with digital transformation strategies will expand external data sources by at least 3- to 5-fold and delivery of data to the market by 100- to 1000-fold or more.

6. By 2018, there will be 22 billion internet of things devices installed, driving the development of over 200,000 new internet of things apps and services.

7. By 2018, over 50% of developer teams will embed cognitive services in their apps (vs. 1 percent today), providing U.S. enterprises with over $60 billion annual savings by 2020.

8. By 2018, over 50% of enterprises will create and/or partner with industry cloud platforms to distribute their own innovations and source others’.

9. By 2018, 80% of B2C and 60% of B2B enterprises will overhaul their “digital front door” to support 1,000 to 10,000 times more customers and customer touch points.

10. By 2020, more than 30% of the IT vendors will not exist as we know them today, requiring realignment of preferred vendor relationships.

“By 2018, at least half of IT spending will be cloud-based, ...”

“By 2017, over 50% of IT spending will be for third platform technologies (cloud, mobile, social business and big-data analytics)...”

IBM Today ...

Applications

Solutions

Business Transformation

Platform

Connecting data

Powered by IBM Watson
The Birth of Watson ...
Watson today ...

The Watson that competed on Jeopardy! in 2011 comprised what is now a single API—Q&A—built on five underlying technologies. Since then, Watson has grown to a family of 28 APIs. By the end of 2016, there will be nearly 50 Watson APIs—with more added every year.
Dark Data

USED DATA

DARK DATA
sensor data that is never utilised

90%

45% no analysis tools setup
28% only analyzes structured data
17% Tools can’t make sense of data

9/12/2016
IBM Research - Zurich Lab
Rethinking Sustainable Cities
IoT Monitoring on IBM's Cloud Platform

Machine Learning

Dashboard

Complex Events Processing

Action Engine

Requirements, Models

IoT Foundation

Data Preparation

Algorithms

Real Time

Off-line

DATA
Rethinking Wellbeing

Tracking steps has been shown to increase daily activity by an average of 2,000 steps.

Did you know? There are 45 million males in our community in your age group (31-35 years old).
We’re already changing the world ...

350+ Watson ecosystem collaborators

750 IoT patents, three times more than any other company

4,000 IoT clients, including leaders in a diverse set of global industries

8,000 new IBM Bluemix® platform users per week

26 billion daily inquiries into The Weather Company’s real-time, mobile-enabled IoT platform

77,000+ developers globally using IBM Watson Developer Cloud services

$3 billion IBM’s four-year investment in cognitive IoT, Weather Company, new HQ in Munich
Hardware in the World of Cloud and Cognitive Computing

New Technologies @
- Device
- Interconnect
- Gate / Macro
- Chip Architecture
- Algorithms

Microprocessor Trends

- Single Thread
- Multi-Core
- Hybrid
- Special Purpose

More active transistors, higher frequency

2005 2015(?) 2025(??) 2035(???)
CPUs are dominating the Cloud ...

... but one size doesn’t fit all and ...

- GPUs boost integer and/or floating point performance
- FPGAs / ASIC can address the performance bottlenecks for
  - complex control flows
  - dataflow computing
  - limited memory capacity
  - memory latency issues
Cognitive Computing Workloads

Towards Cognitive Computing

- Cost
  - Commodity Cluster w/ Hadoop: $O(N)$
  - HPC: $O(N^2)$
  - COMP: $O(N^3)$

- Workloads:
  - Simple DB queries
  - Information Retrieval
  - Clustering
  - Graph Analytics
  - Knowledge Graph Creation
  - Dim. Reduction
  - Uncertainty Quantification
Inter-node vs. Intra-node Heterogeneous Computing Systems

- **hadoop-style workloads**
  
  - **main metrics**
    - cost (capital, energy)
    - compute density
    - scalability
  
  ➔ specialized, homogeneous nodes
  ➔ datacenter disaggregation

- **complex HPC-like workloads**
  
  - **main metrics**
    - memory / accelerator / inter-node BW
    - data centric design
    - heterogeneous compute resources
  
  ➔ versatile, heterogeneous nodes
Heterogeneous Nodes: POWER8 Accelerator Interfaces

- GPU
- POWER8+ Processor
- IBM & Partner Devices
- Memory Interface Control
- Server Class Memory

Connections:
- NVLINK
- CAPI
- 8 x 28.8 GB/s DMI
CAPI ... Coherent Accelerator Processor Interface

Standard I/O Model Flow

DD Call → Copy/Pin → MMIO Notify → Accelerate → Poll / Int → Copy/Unpin → Return DD

Flow with a Coherent Model

Shared Mem. Notify Accelerator → Accelerate → Shared Memory Completion

POWER8 Processor

CAPI FPGA

POWER Service Layer

AFU 0

AFU 1

AFU 2

AFU n
Accelerated Fast Fourier Transformation Library

FFTs are widely used in cognitive computing ...

- Data preparation: spectral analysis, filter banks
- Data compression: MP3, JPEG
- ML: convolutional neural networks [1]
- HPC: partial differential equations, mathematical finance

Common FFT Libraries (FFTW, ESSL, MKL,…)

FFT W on Heterogeneous Compute Nodes

User GNURadio model

GNURadio (dynamically linking fftw)

User application

cuFFT library  FFTW library  Custom FFT API  Custom FFT API

CUDA runtime  User mode driver

CUDA driver  Device driver  libcxl

NVIDIA K80  POWER8 CPU  PCIe FPGA  CAPI FPGA

POWER system

select optimal platform here

train mapping strategy using sensors

Performance

Power

http://openpowerfoundation.org/presentations/energy-efficient-transparent-library-acceleration-with-capi/
... for a single CAPI FFT call is
• 10% higher than CPU (can be improved as the AFU is bandwidth optimized)
• 4x better compared to a PCIe version using OpenCL
Test case: Compute 100 rounds of 32768 subsequent 4k-point FFTs in complex single precision float (1GB input samples per round)

a) 1 core  
   10.6 GFLOP @ 50W  
   = 0.21 GFLOP/W

b) 12 cores  
   33.5 GFLOP @ 108W  
   = 0.31 GFLOP/W

c) 12 cores  
   30.6 GFLOP @ 193W  
   = 0.12 GFLOP/W

d) 1 AFU  
   23.6 GFLOP @ 7W  
   = 3.37 GFLOP/W

e) 1 GPU  
   38.3 GFLOP @ 132W  
   = 0.29 GFLOP/W

1) 12 threads, SMT1, DVFS off
2) 96 threads, SMT8, DVFS on
3) NVIDIA K40, CUDA-7.5

Result: One AFU is 2.2x faster and 16x more energy efficient compared to one core
More Examples?

- **Sparse Matrix Operations ...**
  ... far from peak performance on CPUs and GPUs

- **Stochastic Matrix-Function Estimator (SME)**
Near-memory Acceleration

- big-data analytics, neural networks, cognitive computing, graph algorithms, ... benefit from low latency, small access granularity, and large memories.
- memory performance and power depend on a complex interaction between workload and memory system
  - locality of reference, access patterns/strides, ...
  - cache size, associativity, replacement policy, ...
  - bank interleaving, refresh, row buffer hits, ...
- current systems use “bare metal” programming to adapt workload to memory system
- memory system should be programmable / adaptive
- must integrate programmable compute capabilities to achieve substantial performance & power gains for a wide range of workloads

- enabling near-data processing capabilities, while being minimally-invasive, in an existing CPU architecture

- ability to implement wide range of near-data processing functionality from optimized fixed-function hardware to a multiprocessor SOC

- dereferencing all virtual pointers of the host process on the NDP, coherent with the CPUs view of the memory
Near-memory Acceleration Demo: ConTutto

- conTutto replaces memory buffer (Centaur) with an FPGA
- in-system experiments with our near-memory accelerator concept at full speed
- joint work with Yorktown ConTutto team on a generic Accelerator interface
- FFT and other kernels successfully demonstrated
The OpenPOWER Foundation – 200+ Members & Growing
Inter-node vs. Intra-node Heterogeneous Computing Systems

- **hadoop-style workloads**
  - main metrics
    - compute density
    - cost (capital, energy)
    - scalability
  - specialized, homogeneous nodes
  - datacenter disaggregation

- **complex HPC-like workloads**
  - main metrics
    - memory / accelerator / inter-node BW
    - data centric design
    - heterogeneous compute resources
  - versatile, heterogeneous nodes
ZRL “Dome” μServer of Hyperscale DCs

- **Cloud economics**
  - density (>1000 nodes / rack)
  - integrated NICs
  - switch card (backplane, no cables)
  - medium to low-cost compute chips

- **Passive liquid cooling**
  - ultimate density (cooling >70W / node)
  - energy re-use

- **Built to integrate heterogeneous resources**
  - CPUs
  - Accelerators
CloudFPGA: Network-attached FPGAs in Hyperscale DCs

- **Disaggregation of compute resources**
  - FPGAs can be deployed independent of:
    - the # CPUs (respectively servers)
    - the server form factor (which keep on shrinking)
  - FPGAs can be provisioned / rented similar to other cloud compute, storage and network resources

- **Scalability**
  - Users can build SDN fabrics of FPGAs in the cloud
  - FPGAs are promoted to the rank of peer processor (end of slavery)
  - HW-based FPGA-to-FPGA communication provides low latency and high-Tput (RDMA NICs)
From a practical point of view ...

- A stand-alone appliance/accelerator equipped with an FPGA, (optional) local memory and an integrated network controller interface (iNIC)
- The iNIC enables the FPGA to hook itself to the network and to communicate with other DC resources, such as servers, disks, I/O and other FPGA appliances
SuperVessel: The OpenPOWER Cloud for Developers and Ecosystem

- SPARK, Symphony
- **Accelerator service**
  - Cloud Data Service
  - IoT application development platform
  - POWER open source migration service
  - Machine learning & deep learning
  - Science computation platform

[www.ptopenlab.com](http://www.ptopenlab.com)
Accelerator DevOps Service on OpenPOWER cloud

- Online Accelerator project management
- Online development service with Cloud-based IDE
- Test in VM/Docker equipped with FPGA (for POWER8 & CAPI)
- Publish to Accelerator App. Store and deployment for application on cloud

(Collaboration with Xilinx)

FPGA resource virtualization with Docker
Accelerator scheduling for FPGA resource in Cloud
Data synchronization in DevOps environment
Applications
... demos for new clients to try applications with accelerators.

Accelerators
... allow accelerator developers to create new accelerator and publish it.
... allow application developers to create VM/dockers with the selected accelerators.
Conclusions

- IT industry is going through a phase of transformation (... & IBM, too)
  - cloud is the center of gravity
  - many opportunities, eg, cognitive IoT
- Heterogeneous computing systems are the only sustainable way to advance the two main cloud metrics: € to solution, Time to solution
  - reconfigurable computing is one of the few options available (... In the short term)
  - powerful heterogeneous compute nodes for complex workloads (strong, HPC-like nodes)
    - openpower.org
  - specialized nodes to build rack-level heterogenous systems for hadoop-like applications (eg, cloudFPGA)
- (Hyperscale) Cloud-deployment of heterogeneous computing systems (IaaS) ...
  - is still at the research stage but advancing quickly
    - Supervessel @ www.ptopenlab.com
    - Zurich Heterogeneous Computing Cloud (ZHC2) @ zhc2.zurich.ihost.com
- FPGAs are getting there but standardization & community effort required for
  - accelerator interfaces
  - FPGA compatibility and legacy code
  - cloud orchestration
  - libraries, usage models
But be willing to take incremental steps when you can!