FPL 2016
Lausanne, August 31

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

L.Fiorin, F. Abel, E.Vermij, J.Weerasinghe, S.Dragone M.Purandare, R.Polig, J.vanLunteren, H.Giefers, C. Hagleitner

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

IDC @ 2016: “Transform Or Die”

- IDC:
  30% of tod … them
  today, h …
- IDC:
  1/3 of …
  “disrup …
  revenue, p … that
  they will go o …
- Forrester:
  “Lead The Custo … ation”
- Forrester:
  Custo … son and
  digit … market
  ha … o late
- G …
  “a … e
  ec …
- Gartner:
  The future … an create the
  most effective a … ftware solutions

“30% of today’s tech suppliers will not exist as we know them today …”

“1/3 of the top 20 companies in every industry will be “disrupted” over the next 3 years ...”

Where is the IT Industry going ...???

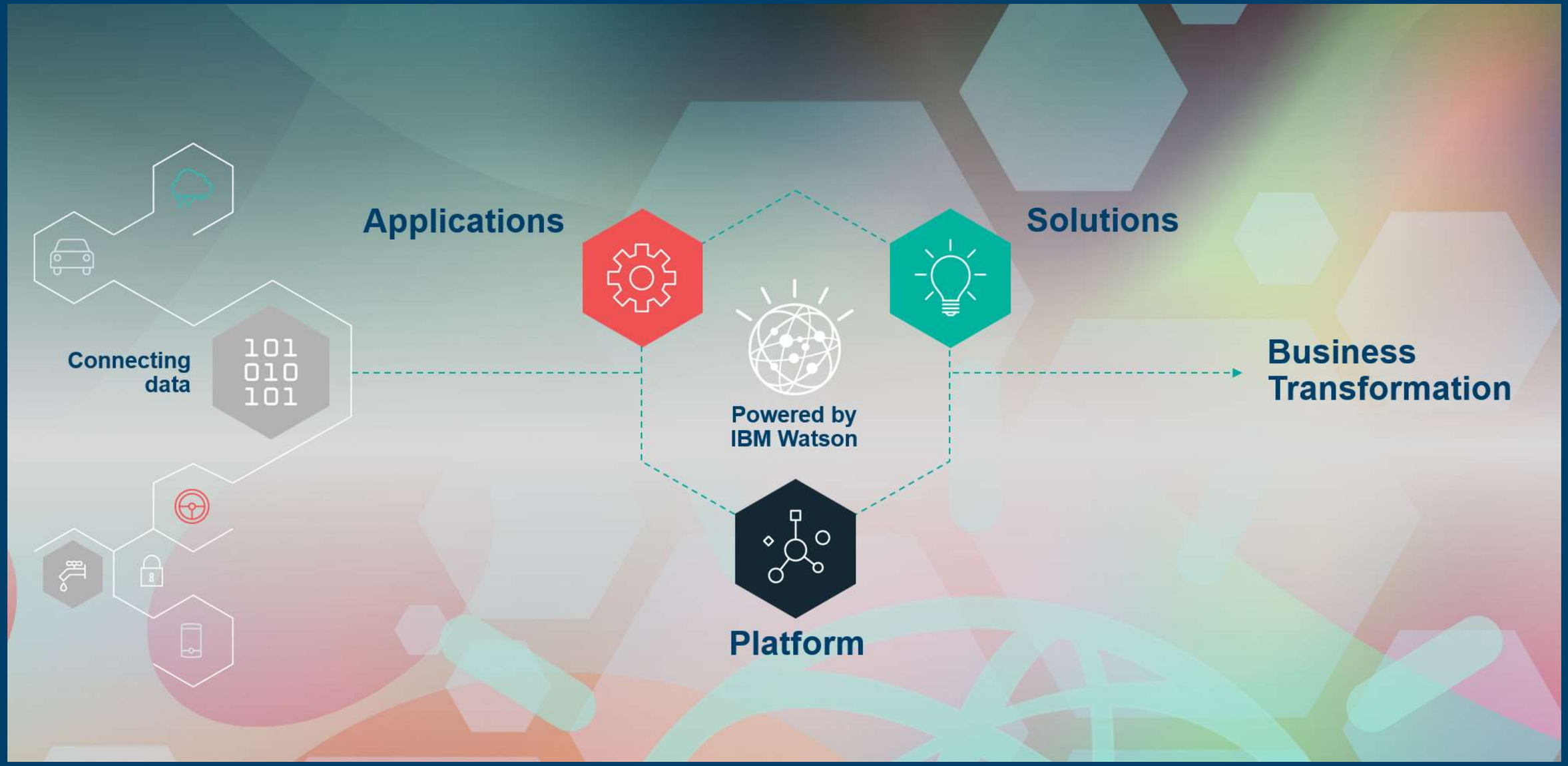
1. By the end of 2017, two thirds of the CEOs of global 2000 enterprises will have **digi** **nter of their corporate strateg**
2. By **ill be for third** and big-data 2020
3. **d**, reaching vices, and
4. egies will two-thirds of apps and se
5. By 201 rategies will expand **external** ry of **data to the**
6. d, **s**
9. **"dig** eir customers and custo
10. By 2020, mor **will not exist as we know them today**, ignment of preferred vendor relationships

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

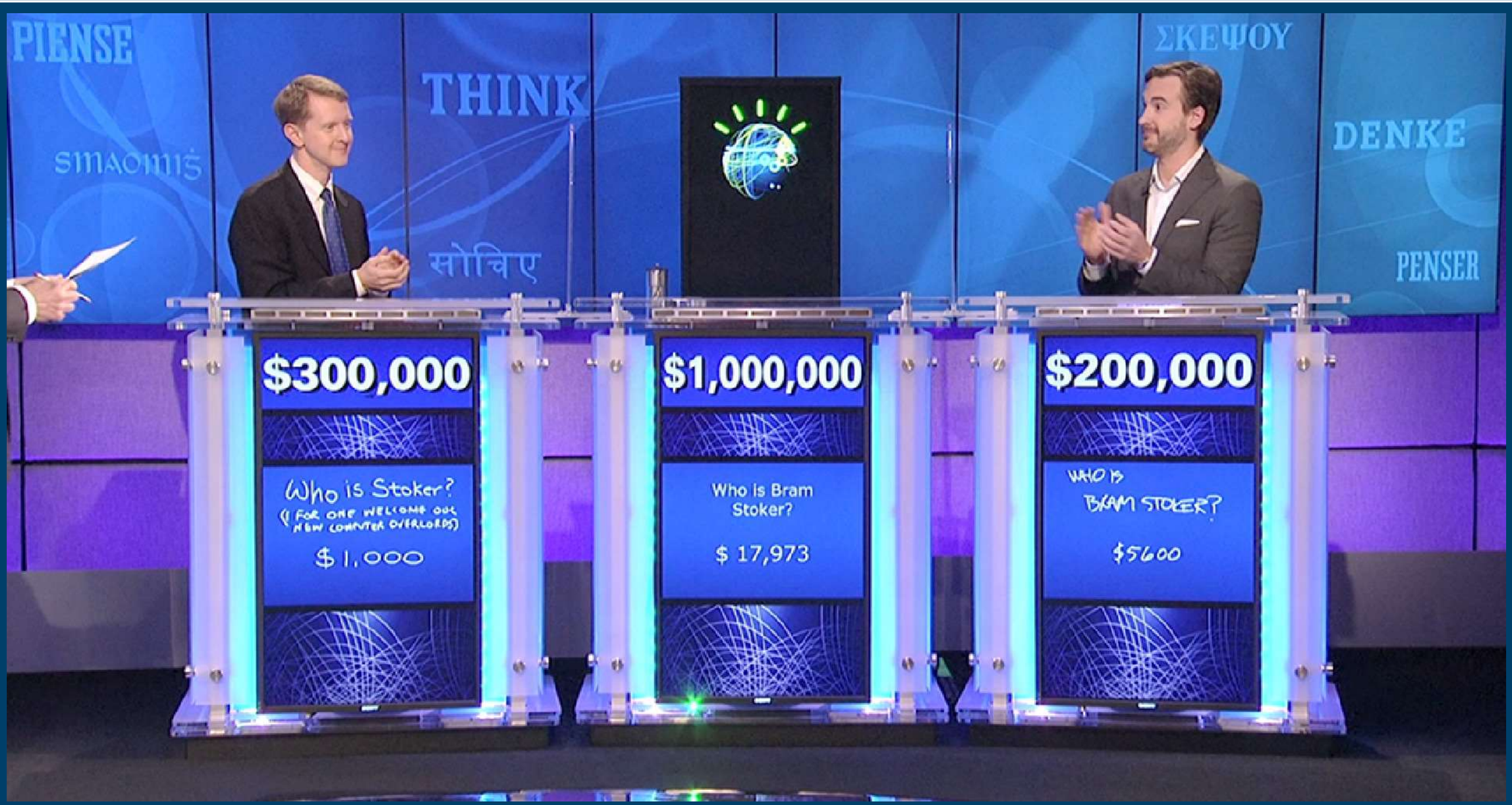
"By 2017,
over 50% of IT spending
will be for third platform technologies
(cloud, mobile, social business and
big-data analytics)..."

IDC FutureScape: Worldwide IT Industry 2016 Predictions — Leading Digital Transformation to Scale

IBM Today ...


Applications
Solutions
Connecting data
101
010
101
Powered by IBM Watson
Business Transformation
Platform

The Birth of Watson ...



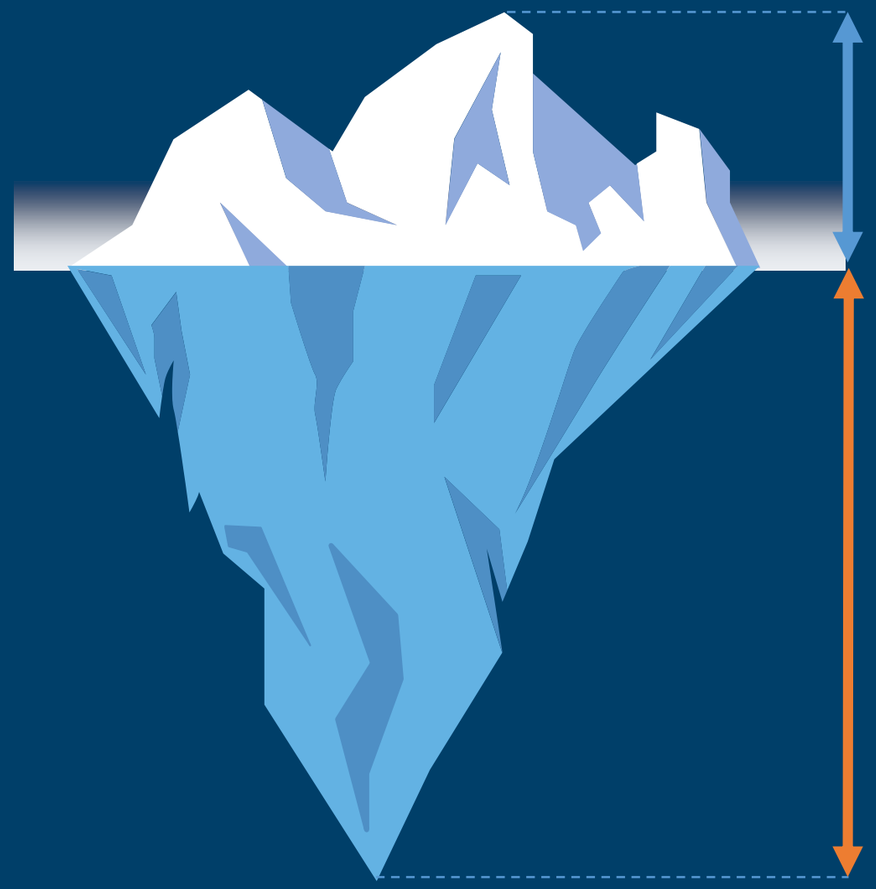
Watson today ...

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**.

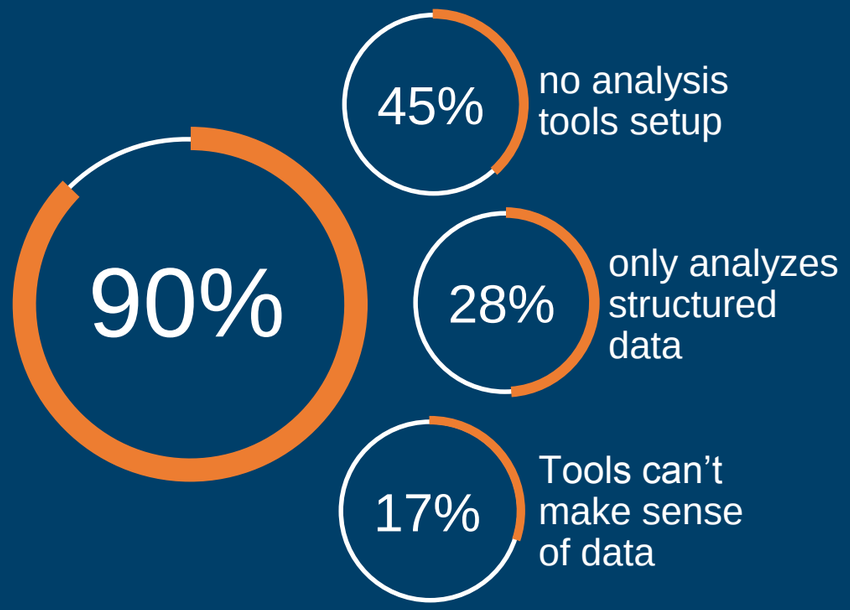
By the end of 2016, there will be nearly **50 Watson APIs**—with more added every year.

Dark Data



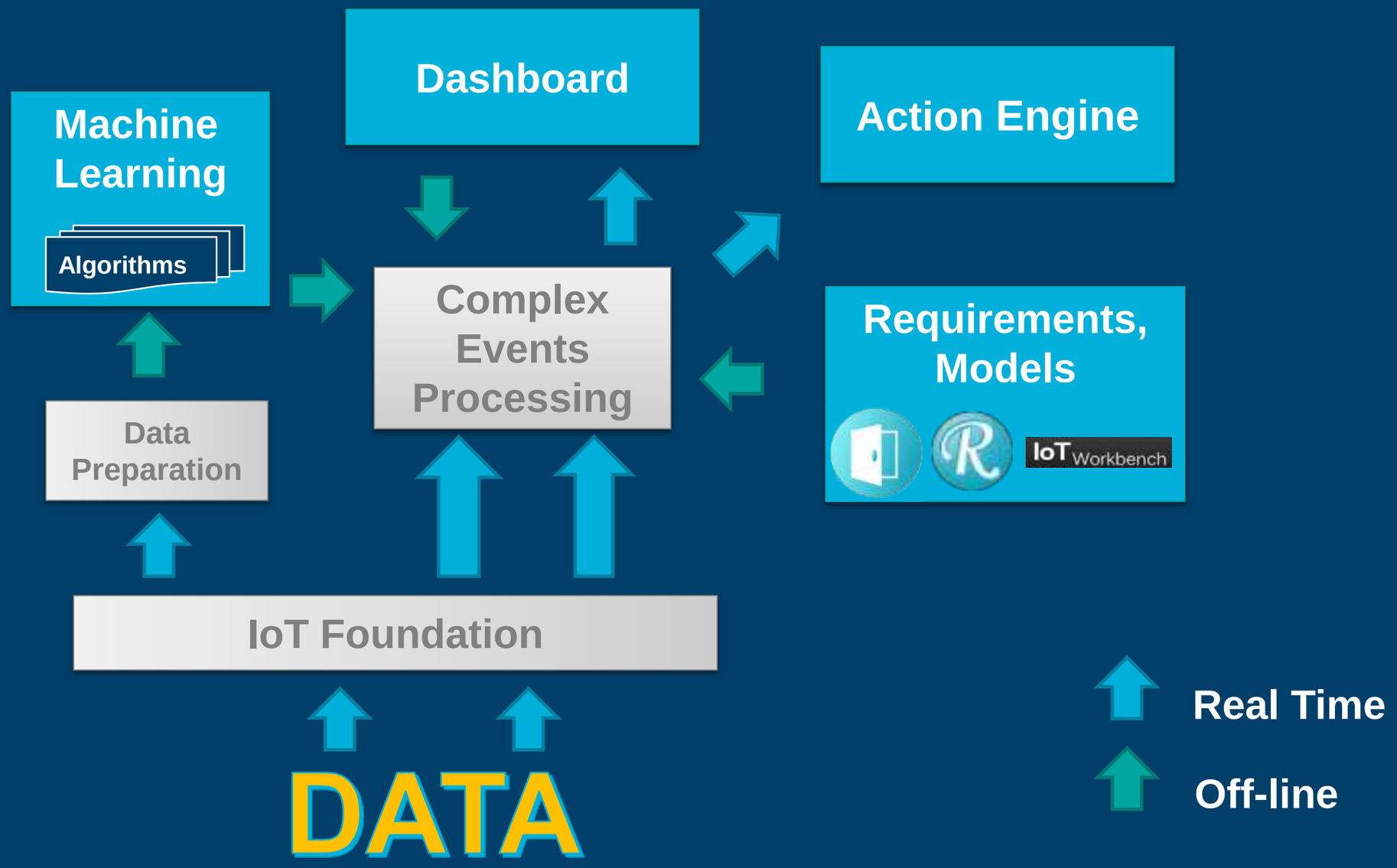
USED DATA

DARK DATA
sensor data that is never utilised

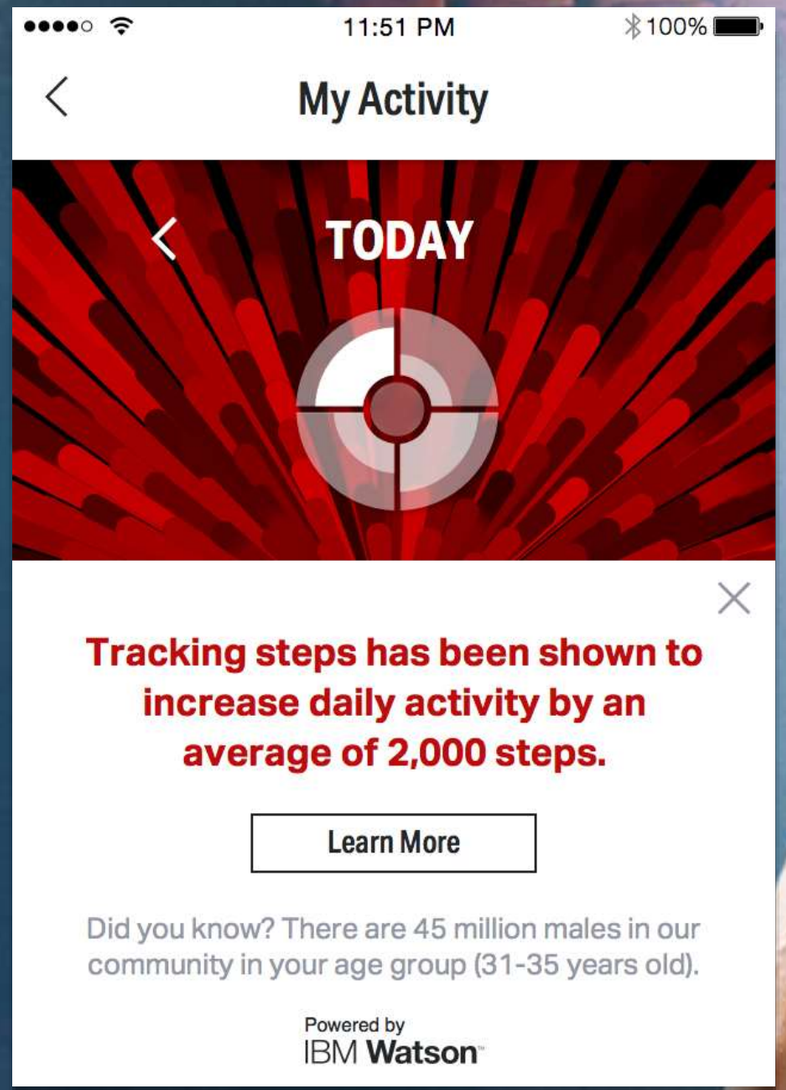


Rethinking Sustainable Cities

IoT Monitoring on IBMs Cloud Platform



Rethinking Wellbeing



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

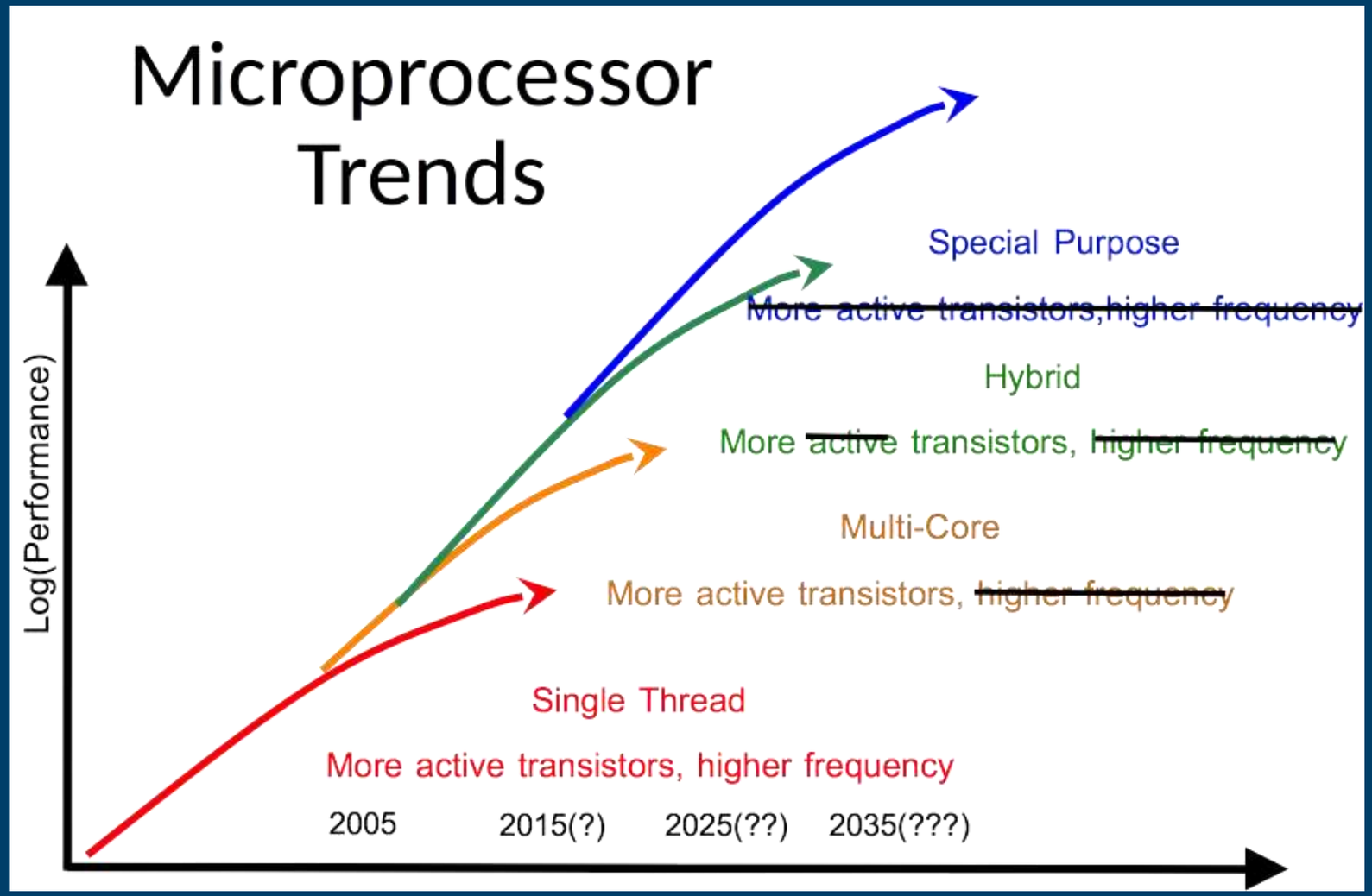
8,000
new IBM Bluemix® platform users per week

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



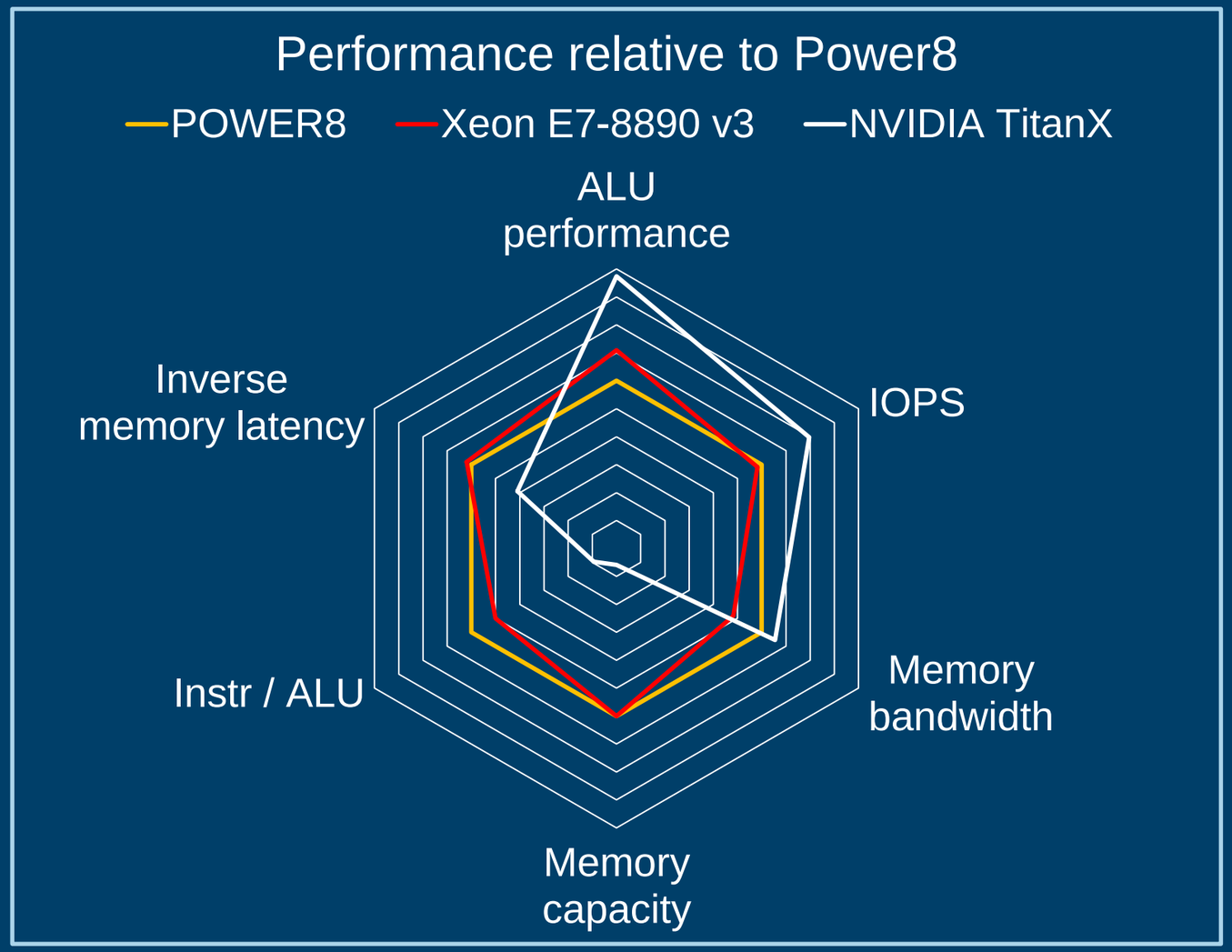
New Technologies @

- Device
- Interconnect
- Gate / Macro
- Chip Architecture
- System Architecture
- Algorithms

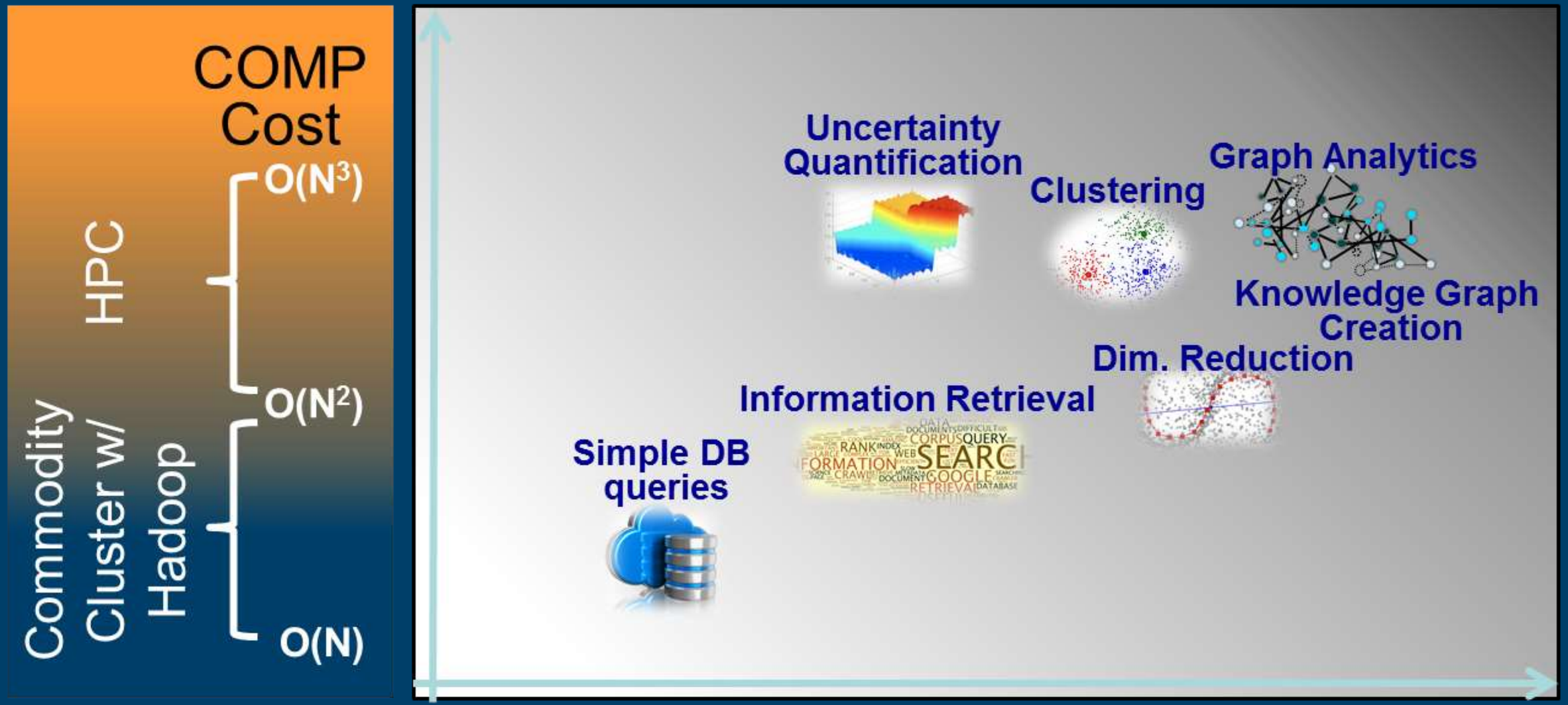
CPUs are dominating the Cloud ...

... 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

Inter-node vs. Intra-node Heterogeneous Computing Systems

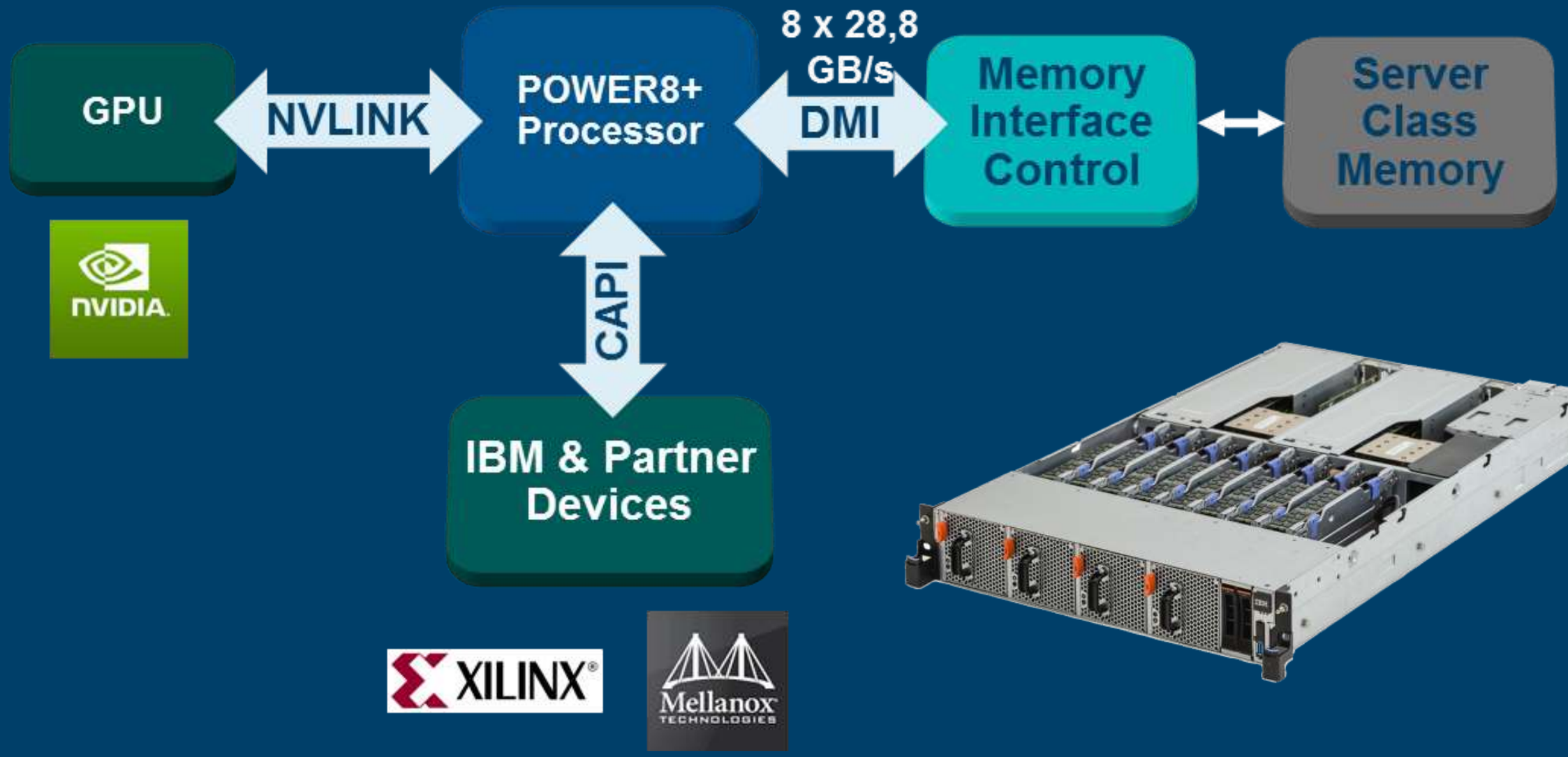
- hadoop-style workloads
- main metrics
  - cost (capital, energy)
  - compute density
  - scalability

→specialiced, 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



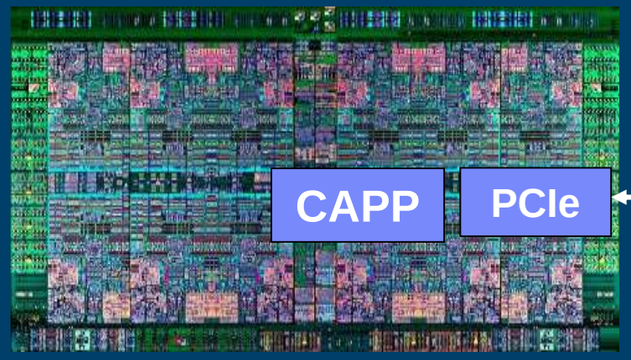
CAPI ... Coherent Accelerator Processor Interface

Standard I/O Model Flow

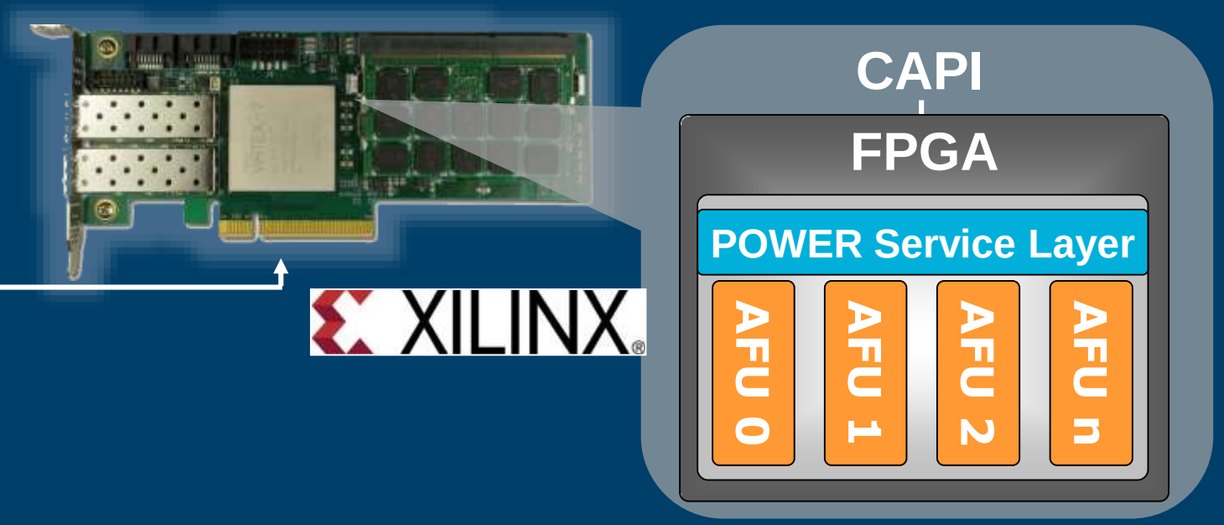
DD Call → Copy/Pin → MMIO Notify → Accelerate → Poll / Int → Copy/Unpin → Return DD

Flow with a Coherent Model

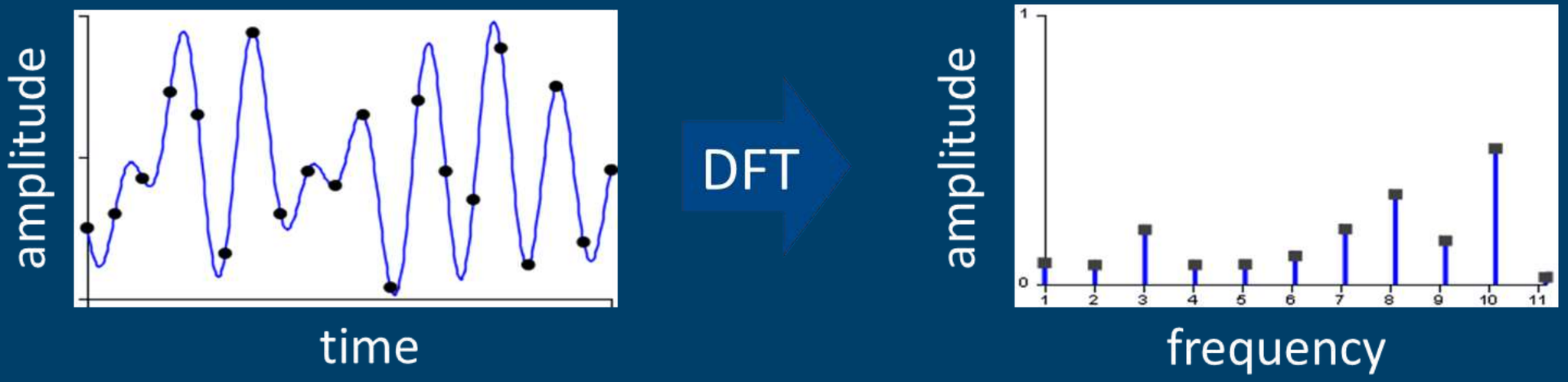
Shared Mem. Notify Accelerator → Accelerate → Shared Memory Completion



POWER8 Processor



Accelerated Fast Fourier Transformation Library



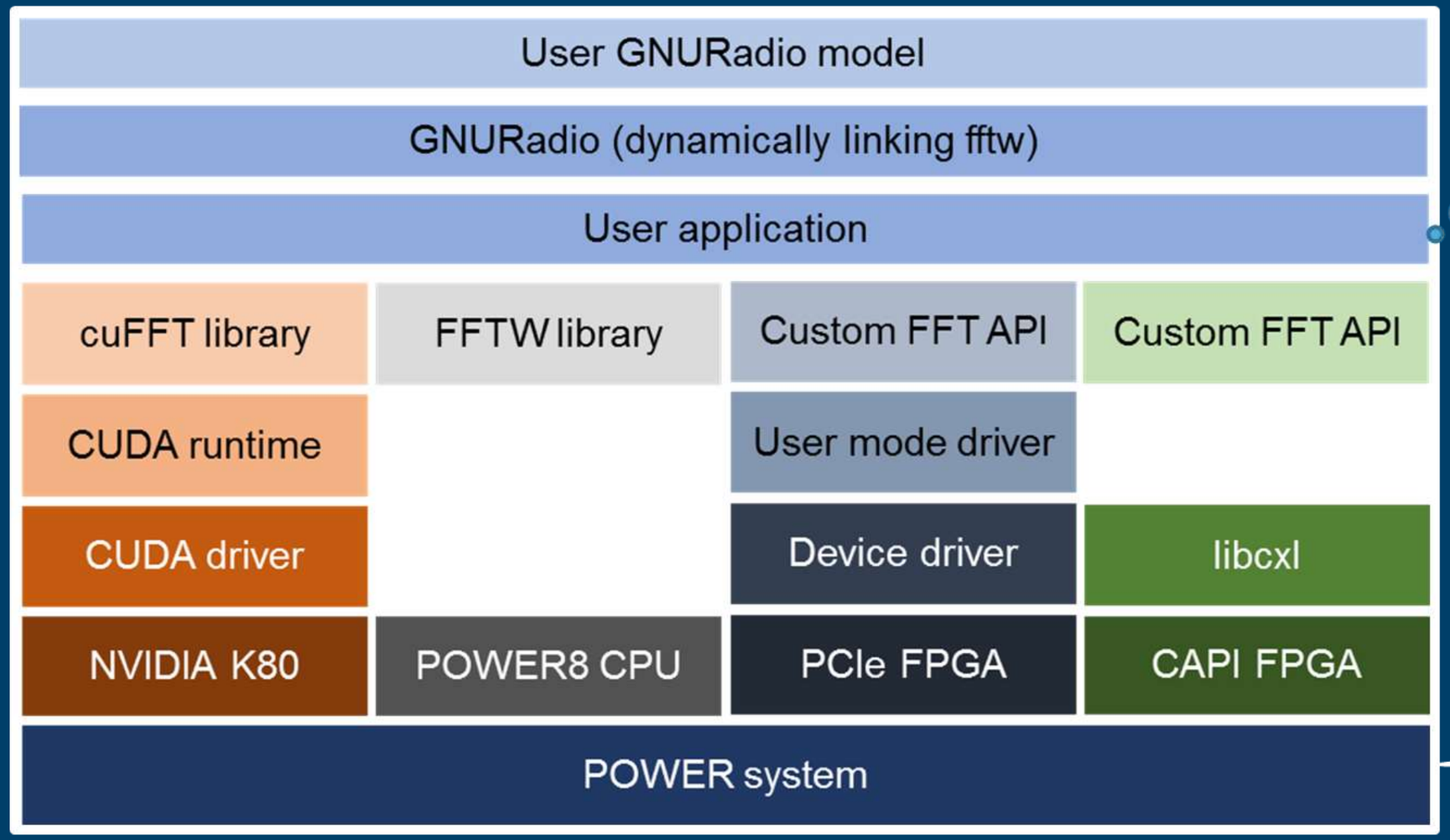
FFTs are widely used in cognitive computing ...

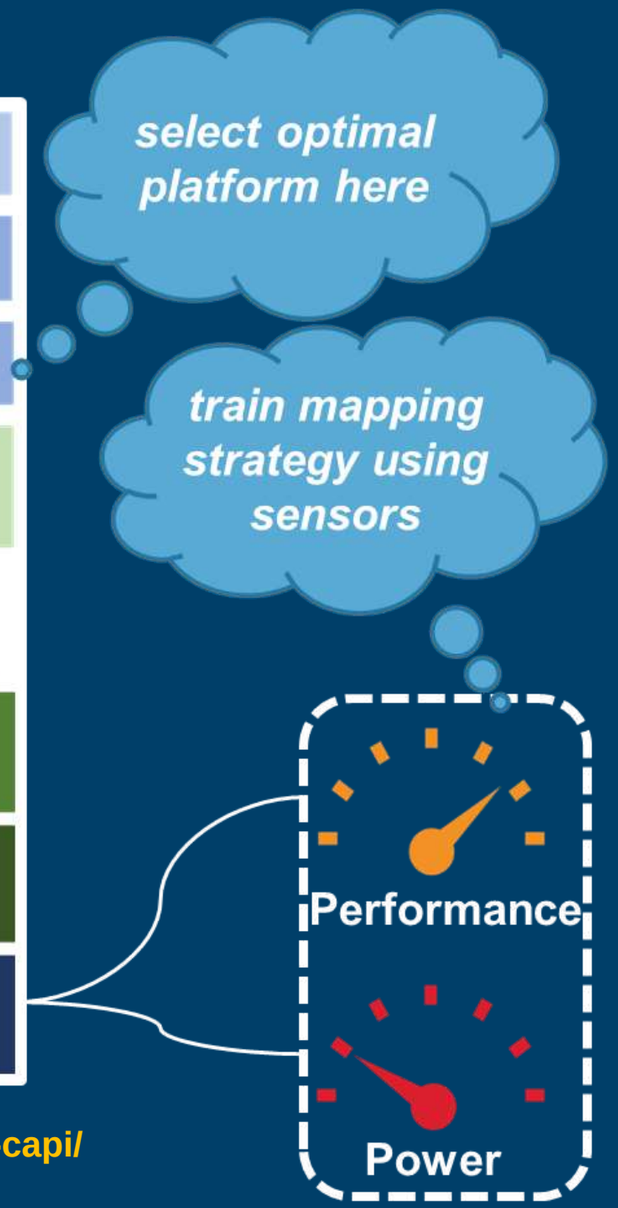
- Data preparation: spectral analysis, filter banks
- Data compression: MP3, JPEG
- ML: convolutional neural networks [1]
- HPC: partial differential equations, mathematical finance

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

[1] Mathieu, Henaff, Lecun. "Fast training of convolutional networks through FFTs". *ICLR'14*

FFTW on Heterogeneous Compute Nodes



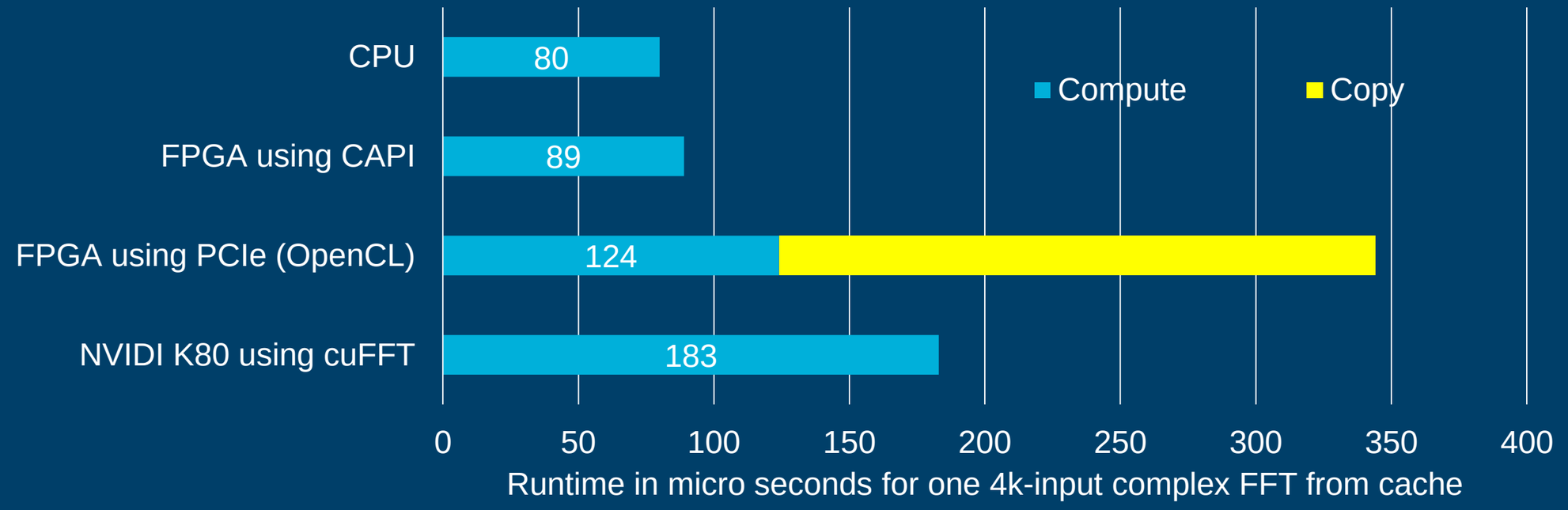


http://openpowerfoundation.org/presentations/energy-efficient-transparent-library-acceleration-with-capi/

Latency

... 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



Performance & Energy Efficiency

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[1)]**	33.5 GFLOP @ 108W	= **0.31 GFLOP/W**
c)	**12 cores[2)]**	30.6 GFLOP @ 193W	= **0.12 GFLOP/W**
d)	**1 AFU**	23.6 GFLOP @ 7W	= **3.37 GFLOP/W**
e)	**1 GPU[3)]**	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

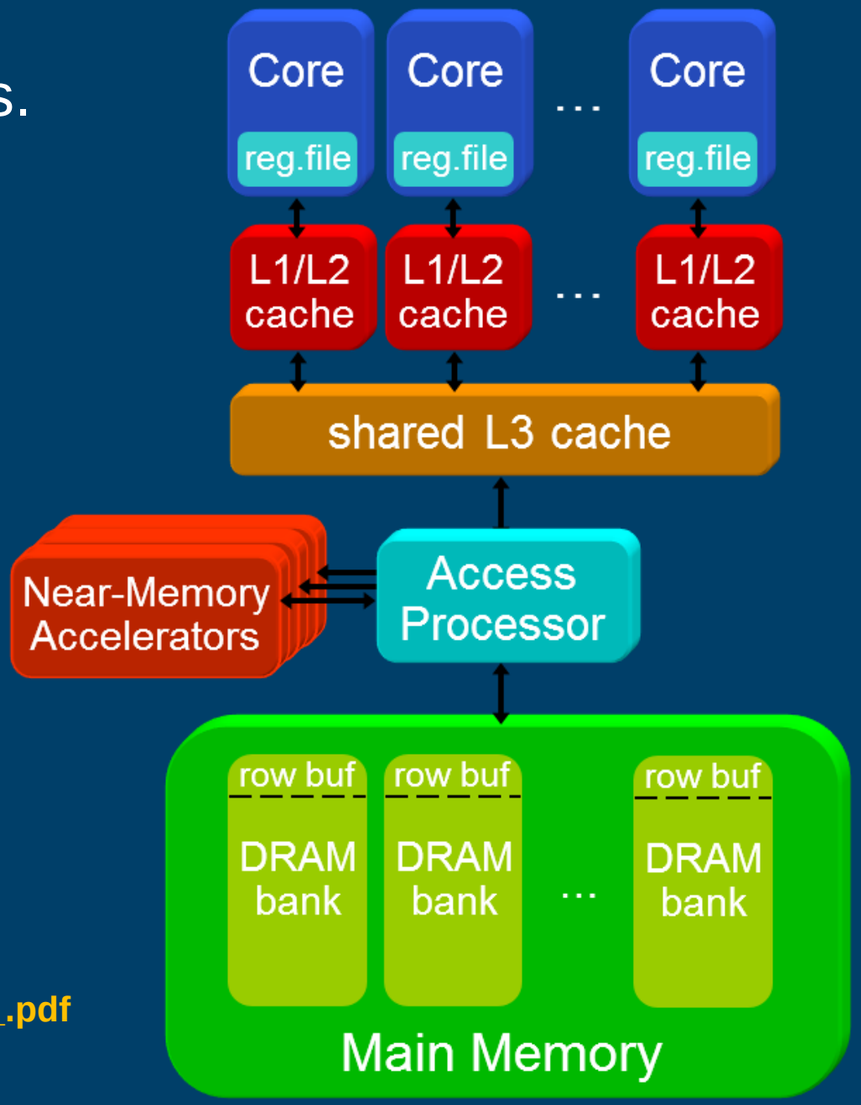
More Examples?

- Sparse Matrix Operations ...
  ... far from peak performance on CPUs and GPUs
  → "Analyzing the Energy-Efficiency of Sparse Matrix Multiplication on Heterogeneous Systems", ISPASS2016
- Stochastic Matrix-Function Estimator (SME)

	Session S5a: Data Analysis and Databases Chair: Sameh Assad *4:00pm–5:30pm* *Auditorium B*
4:00pm	**Energy-Efficient Stochastic Matrix Function Estimator for Graph Analytics on FPGA** Heiner Giefers, Peter Staar, Raphael Polig *IBM Research, Zurich, CH*

Near-memory Acceleration

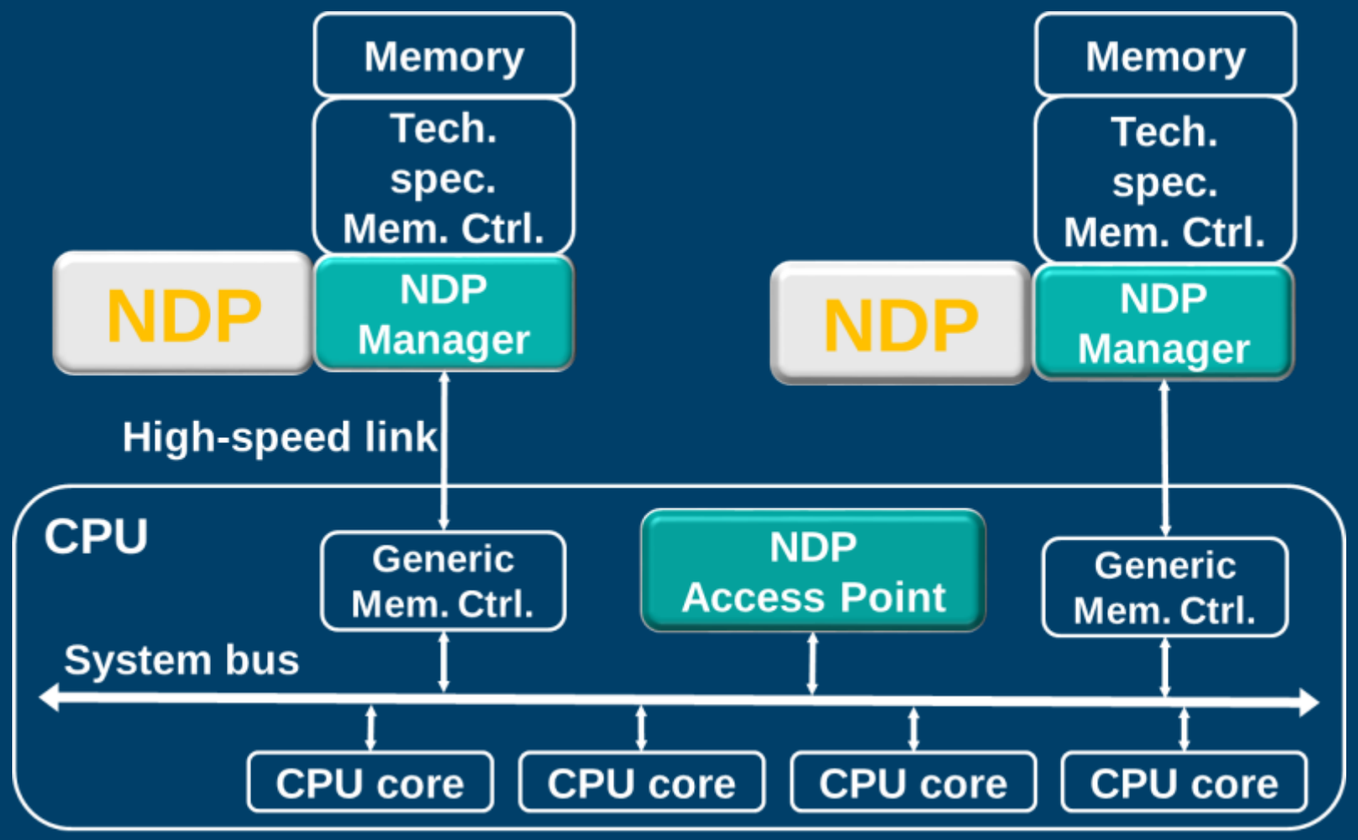
- big-data analytics, neural networks, cognitive computing, graph algorithms, ... benefit from low latency, small access granularity, and large memories.
- memory performance and power depend on a complex interaction between workload and memory system
  - locality of reference, access patterns/strides, ...
  - 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



http://openpowerfoundation.org/wp-content/uploads/2016/03/5_Jan-Van-Lunteren.IBM_.pdf

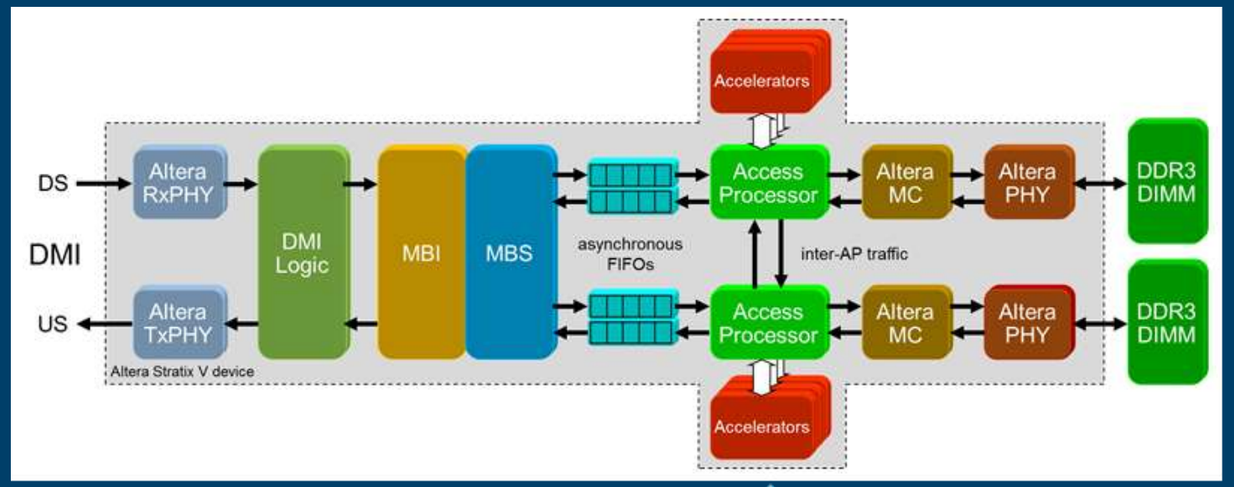
Integrating Near-data Processing in a (POWER) Server

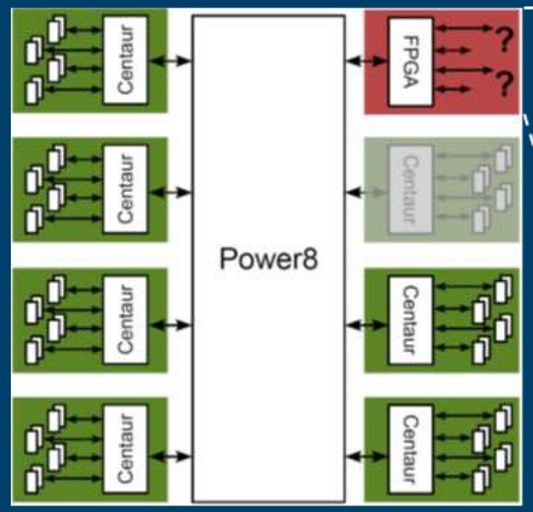
- enabling near-data processing capabilities, while being minimally-invasive, in an existing CPU architecture
- ability to implement wide range of near-data processing functionality from optimized fixed-function hardware to a multiprocessor SOC
- dereferencing all virtual pointers of the host process on the NDP, coherent with the CPUs view of the memory



Near-memory Acceleration Demo: ConTutto

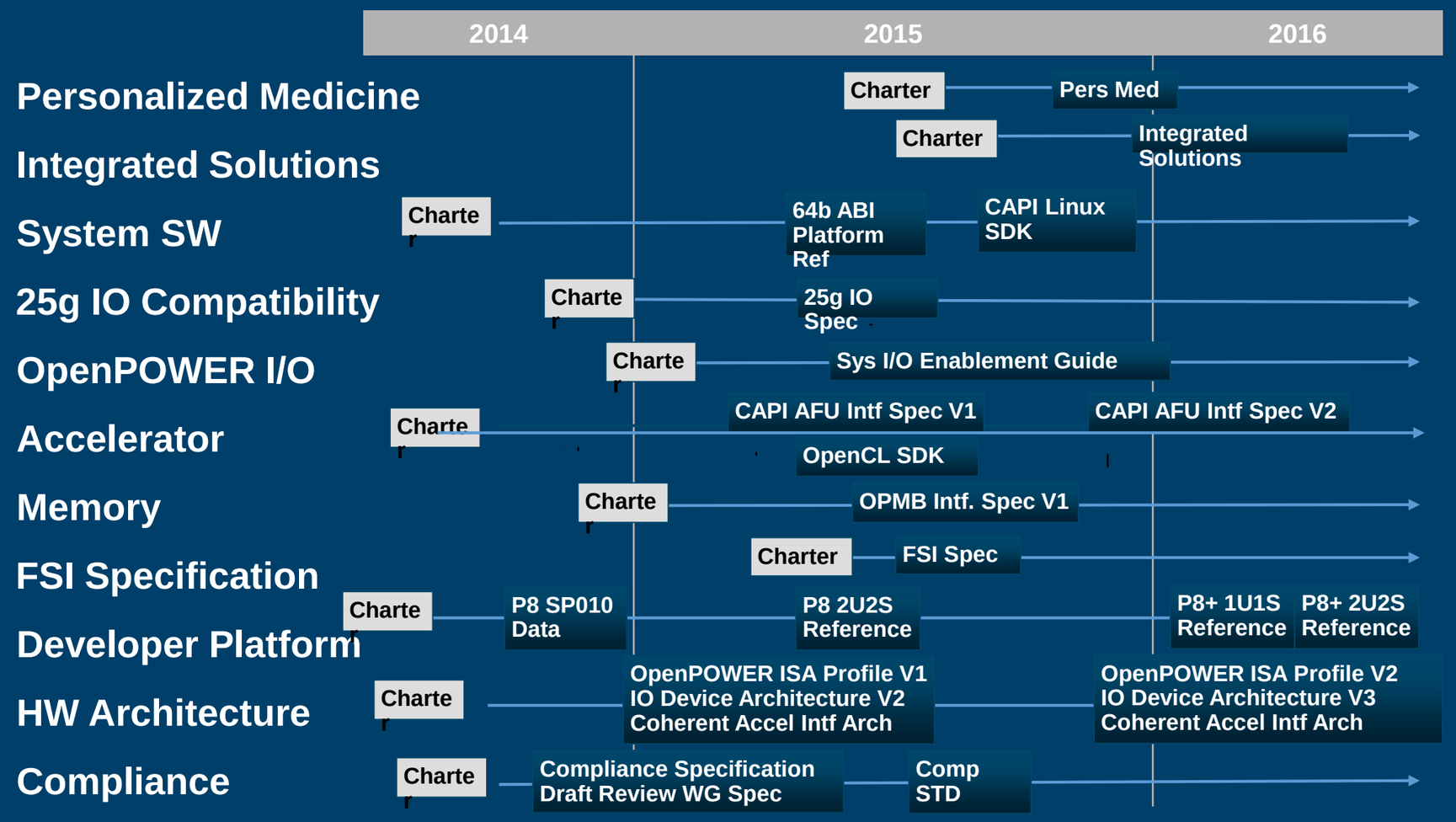
- conTutto replaces memory buffer (Centaur) with an FPGA
- in-system experiments with our near-memory accelerator concept at full speed
- joint work with Yorktown ConTutto team on a generic Accelerator interface
- FFT and other kernels succesfully demonstrated





The OpenPOWER Foundation – 200+ Members & Growing

OpenPOWER Workgroups: Open Standards



SDK – Software Developer Kit

SP010 – Tyan OpenPOWER Customer Reference System

CAPI – Coherent Accelerator Processor Interface

AFU – Accelerator Function Unit

FSI – Field Replaceable Unit (FRU) Service Interface

OPMB – OpenPOWER Memory Bus

ABI – Application Binary Interface

Inter-node vs. Intra-node Heterogeneous Computing Systems

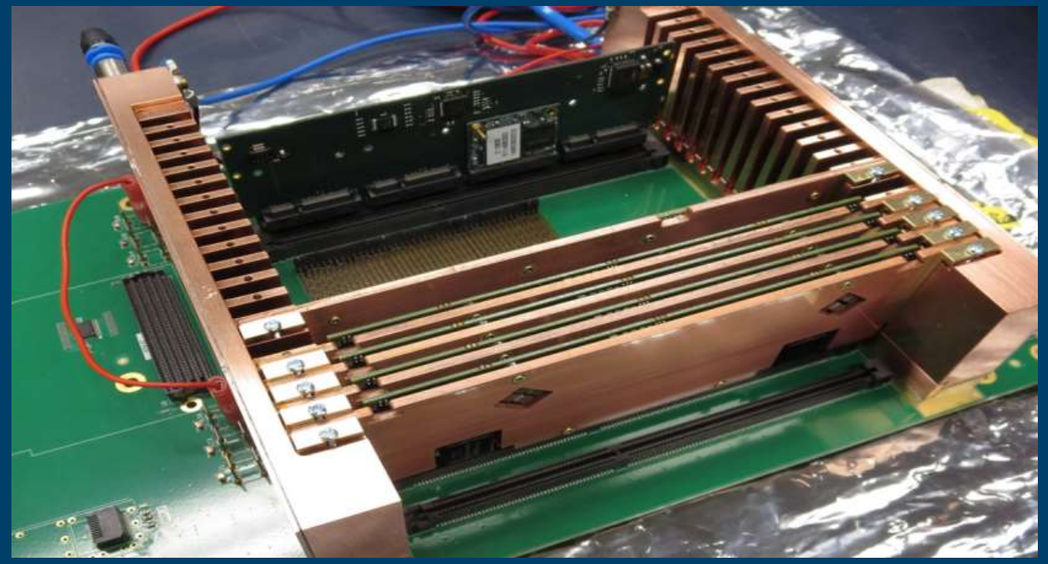
- hadoop-style workloads
- main metrics
  - compute density
  - cost (capital, energy)
  - scalability

→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



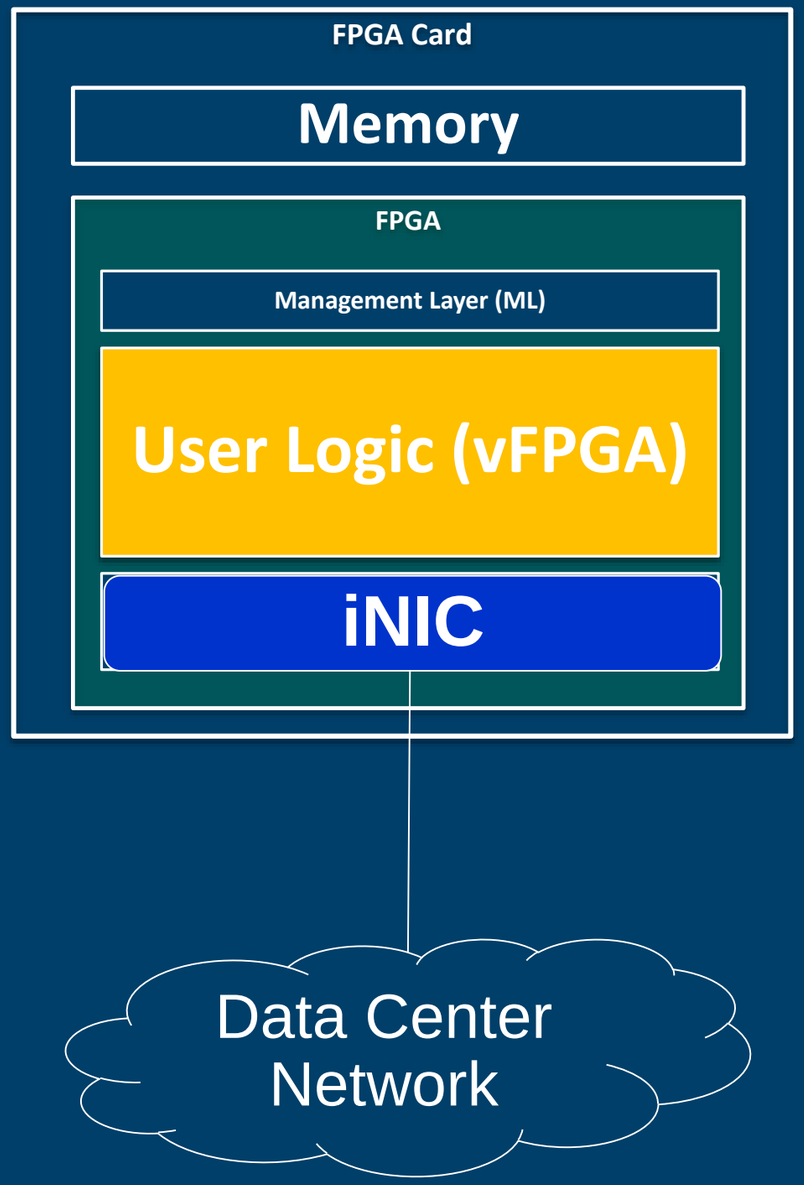
- Cloud economics
  - density (>1000 nodes / rack)
  - integrated NICs
  - switch card (backplane, no cables)
  - medium to low-cost compute chips
- Passive liquid cooling
  - ultimate density (cooling >70W / node)
  - energy re-use
- Built to integrate heterogeneous resources
  - CPUs
  - Accelerators

CloudFPGA: Network-attached FPGAs in Hyperscale DCs

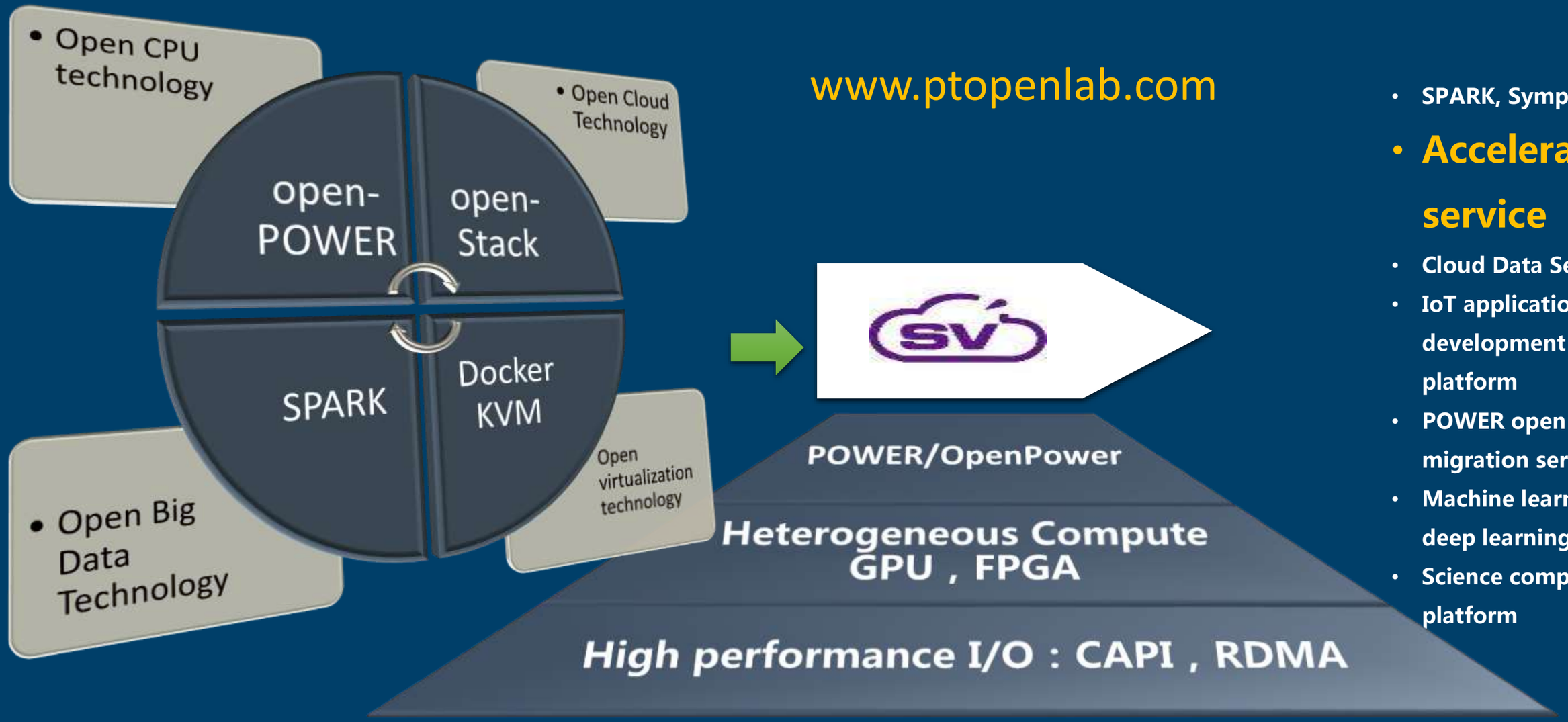
- **Disaggregation of compute resources**
  - FPGAs can be deployed independent of:
    - the # CPUs (respectively servers)
    - the server form factor (which keep on shrinking)
  - FPGAs can be provisioned / rented similar to other cloud compute, storage and network resources
- **Scalability**
  -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



www.ptopenlab.com

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

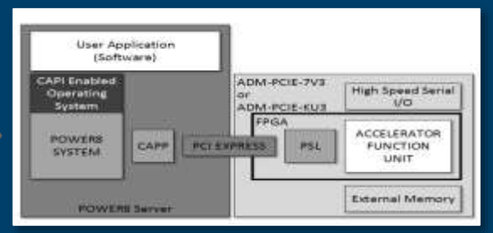
Accelerator DevOps Service on OpenPOWER cloud

Online Accelerator project management

Online development service with Cloud-based IDE

(Collaboration with Xilinx)

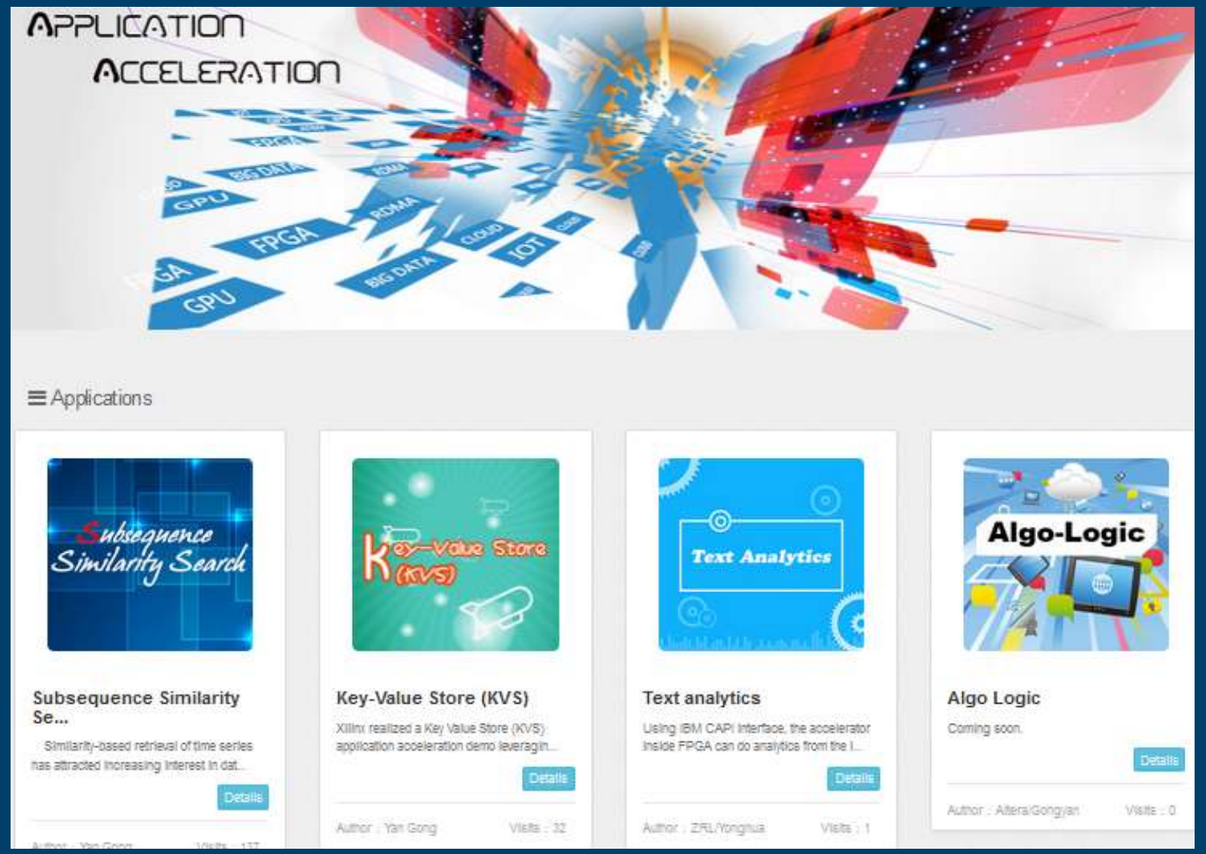
Test in VM/Docker equipped with FPGA (for POWER8 & CAPI)



Publish to Accelerator App. Store and deployment for application on cloud

FPGA resource virtualization with Docker
Accelerator scheduling for FPGA resource in Cloud
Data synchronization in DevOps environment

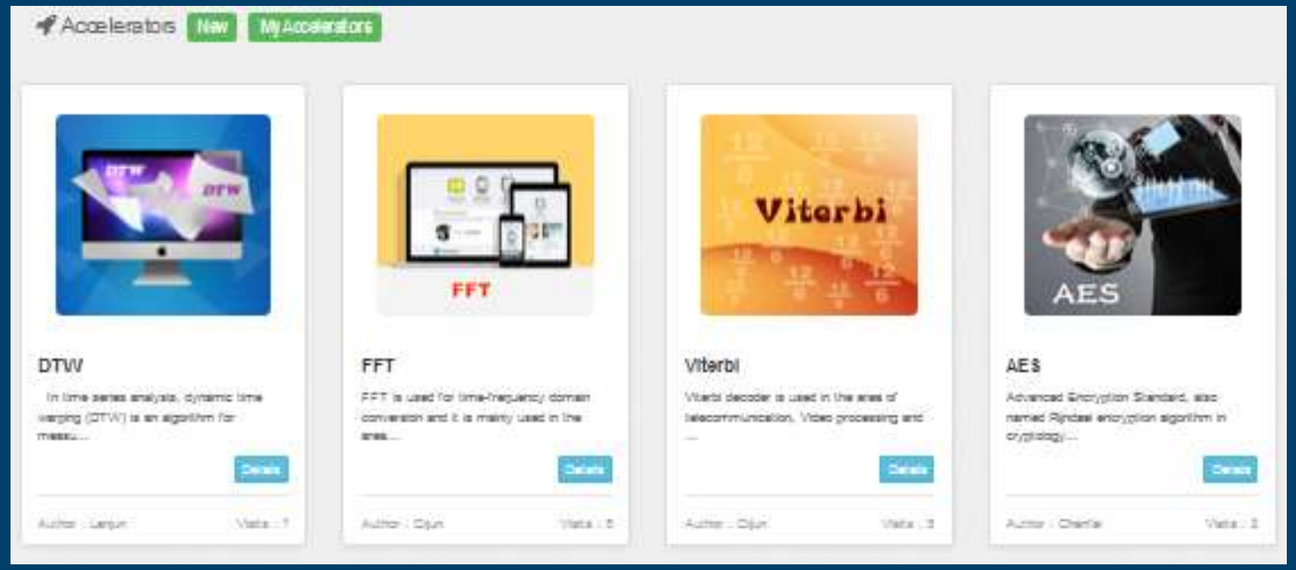
SuperVessel Acceleration App Store



Accelerators

... allow accelerator developers to create new accelerator and publish it.

... allow application developers to create VM/dockers with the selected accelerators



Applications

... demos for new clients to try applications with 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

THINK

BIG