

IBM Data Centric Systems & OpenPOWER

Yoonho Park

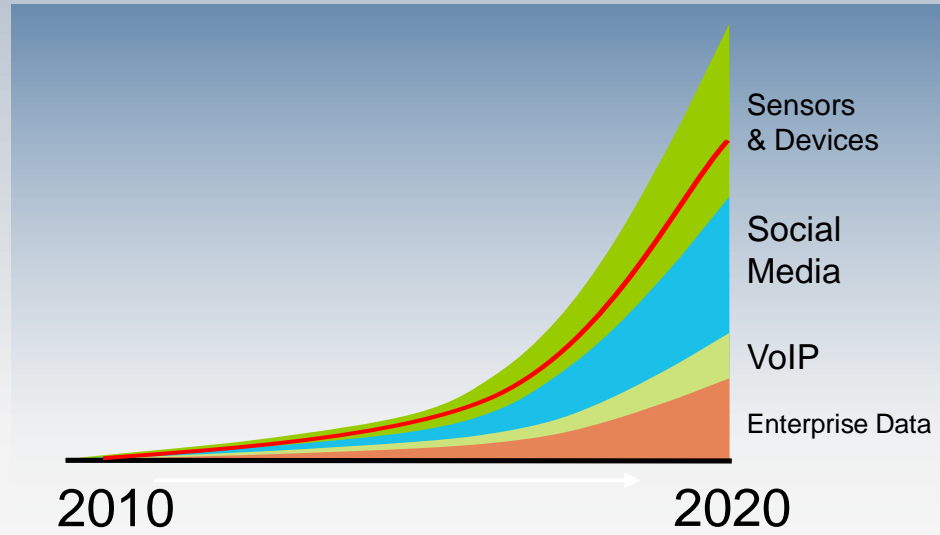
Research Staff Member/Senior Manager

Data Centric Systems Software, Cloud and Cognitive

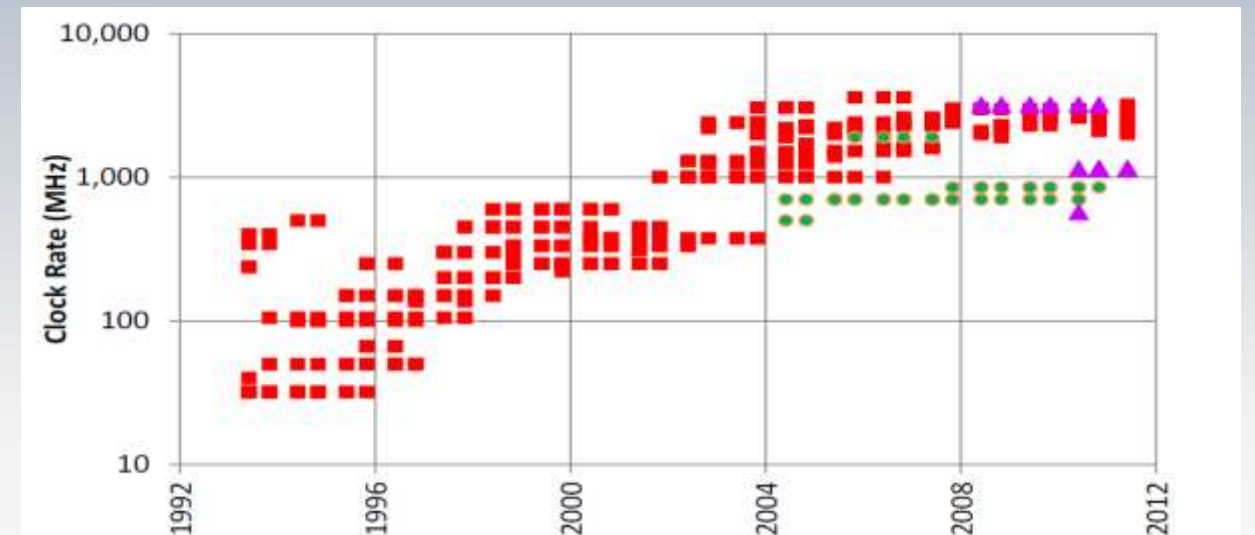
IBM Research

Data Growth Outpaces Computing Technology Elements

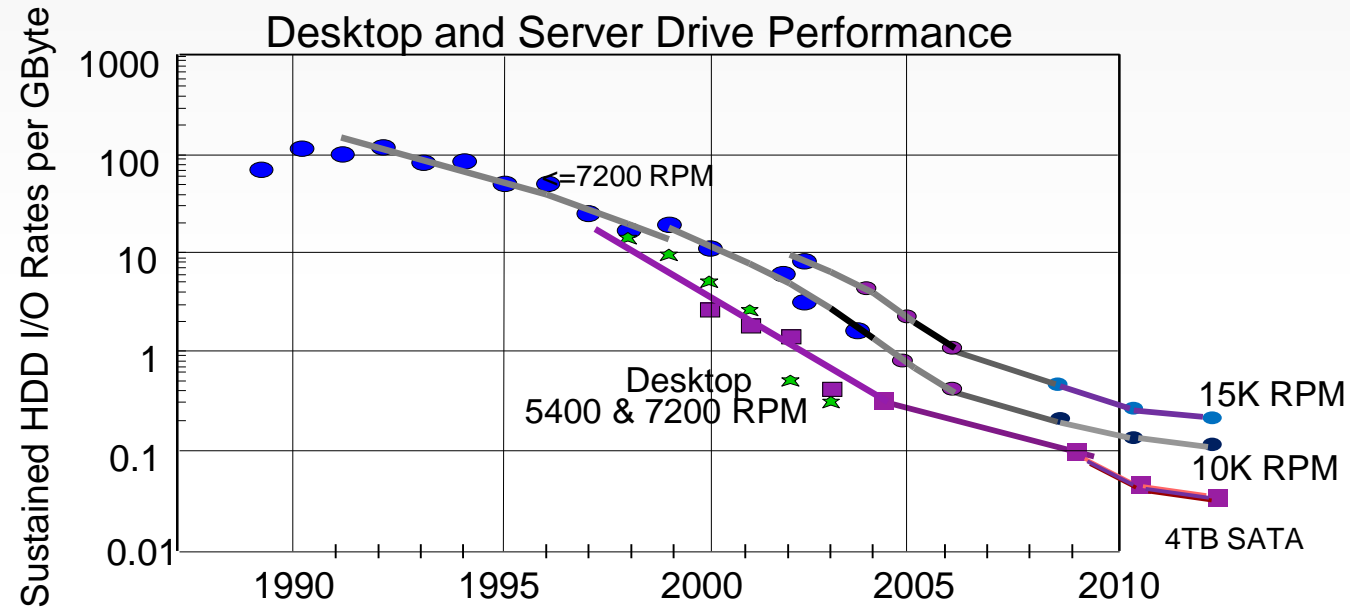
Data volume grows exponentially



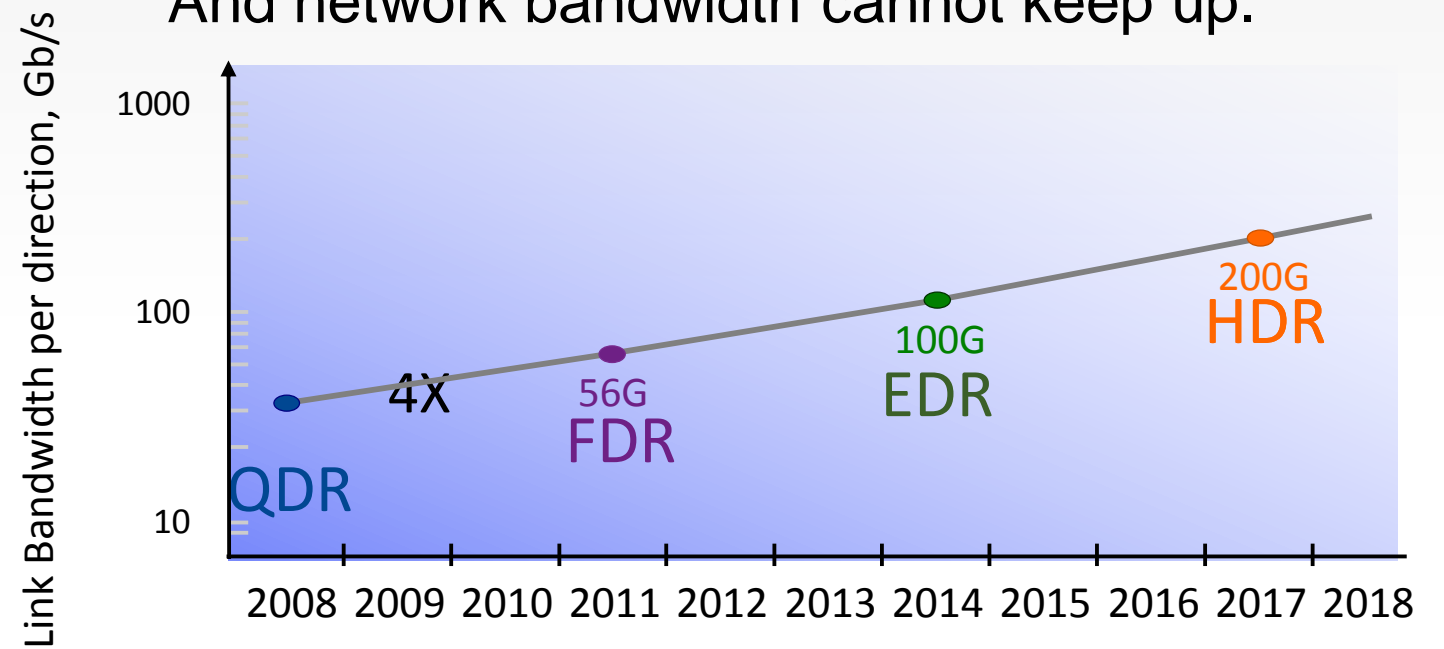
Microprocessor clock rates have stalled...



I/O performance/capacity loosing ground...

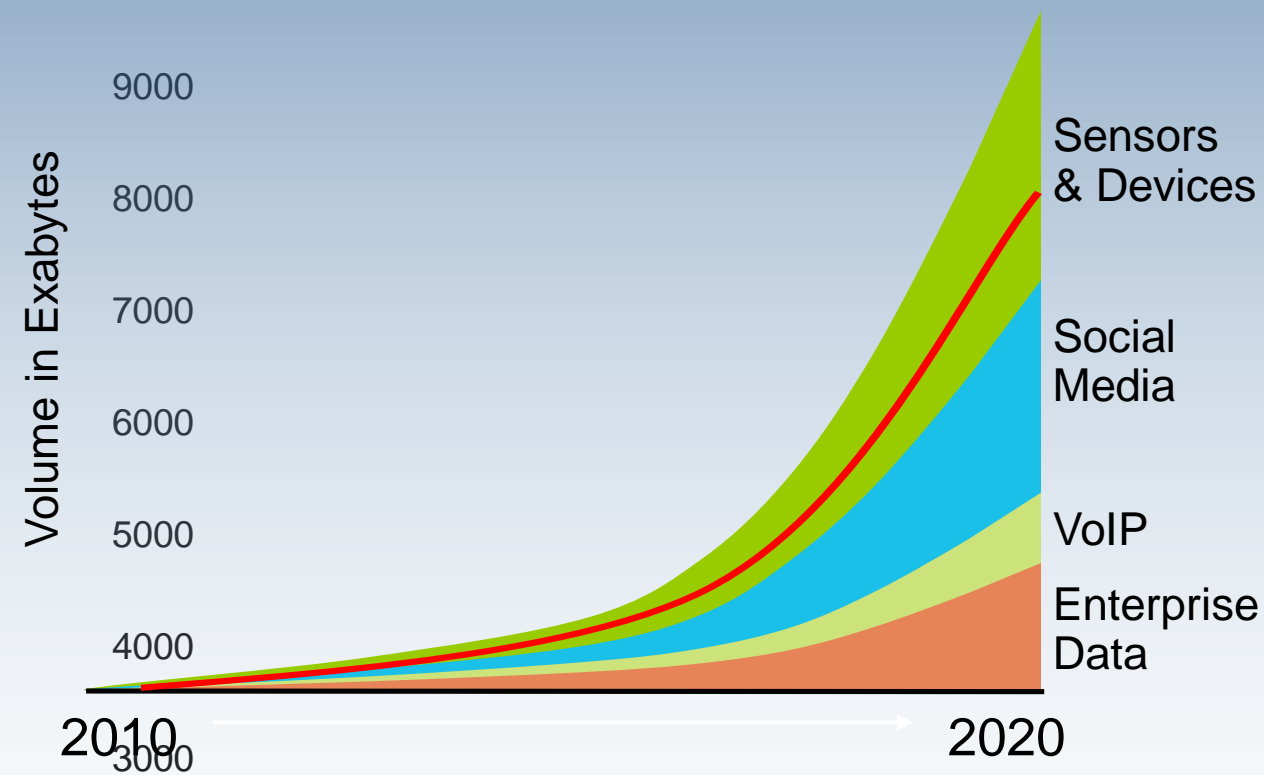


And network bandwidth cannot keep up.



Big Data and the New Era of Computing

Data volume is on the rise



Dimensions of data growth

Terabytes to exabytes of existing data to process

Volume

Streaming data, milliseconds to seconds to respond

Velocity

Structured, unstructured, text, multimedia

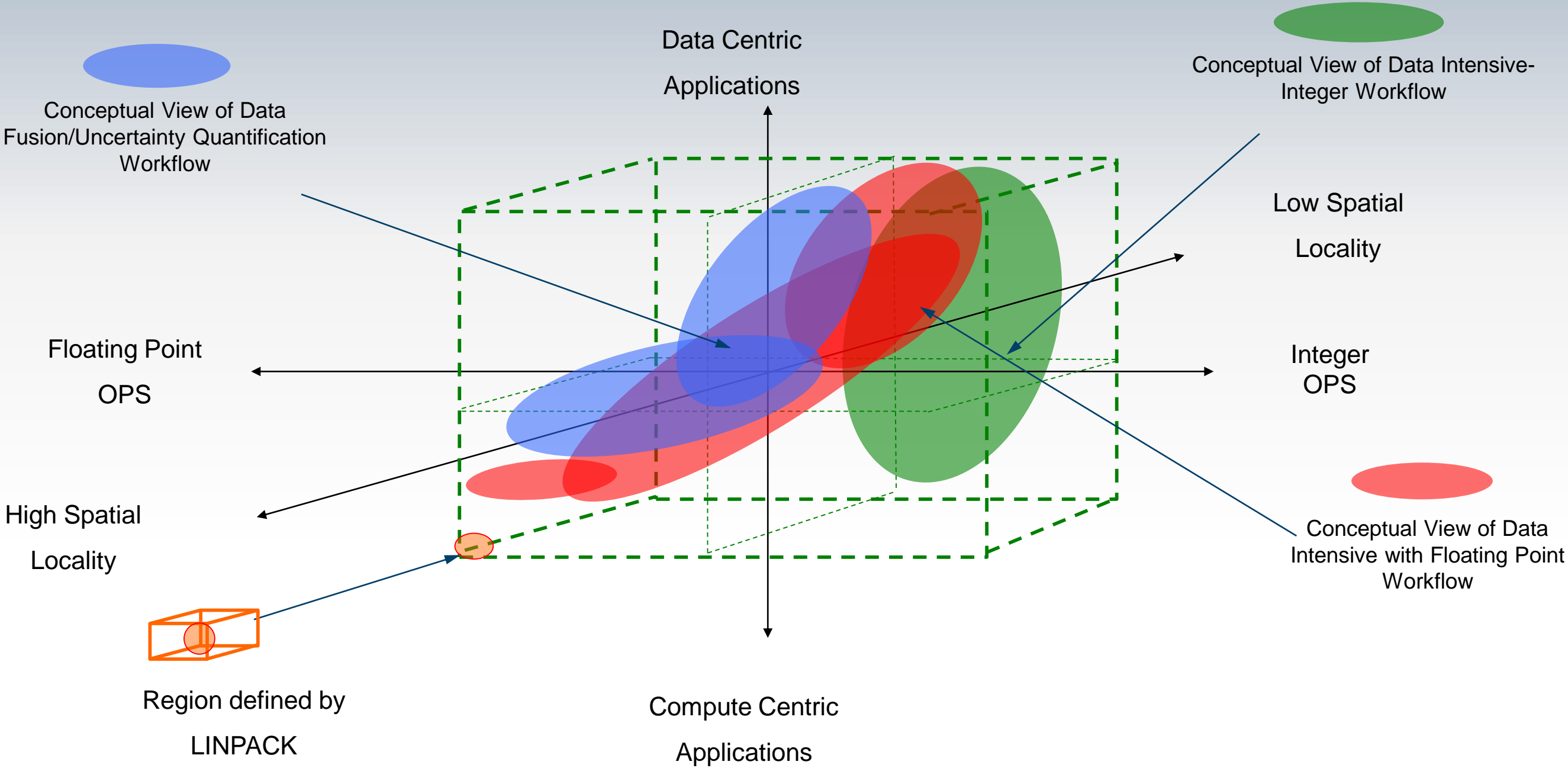
Variety

Veracity

Uncertainty from inconsistency, ambiguities, etc.

- *Big Data analytics and Exascale High Performance Computing facing similar challenges: scale, performance, bandwidth, computational complexity*
- *IBM approach: Move compute to data – Data Centric Systems (DCS)*

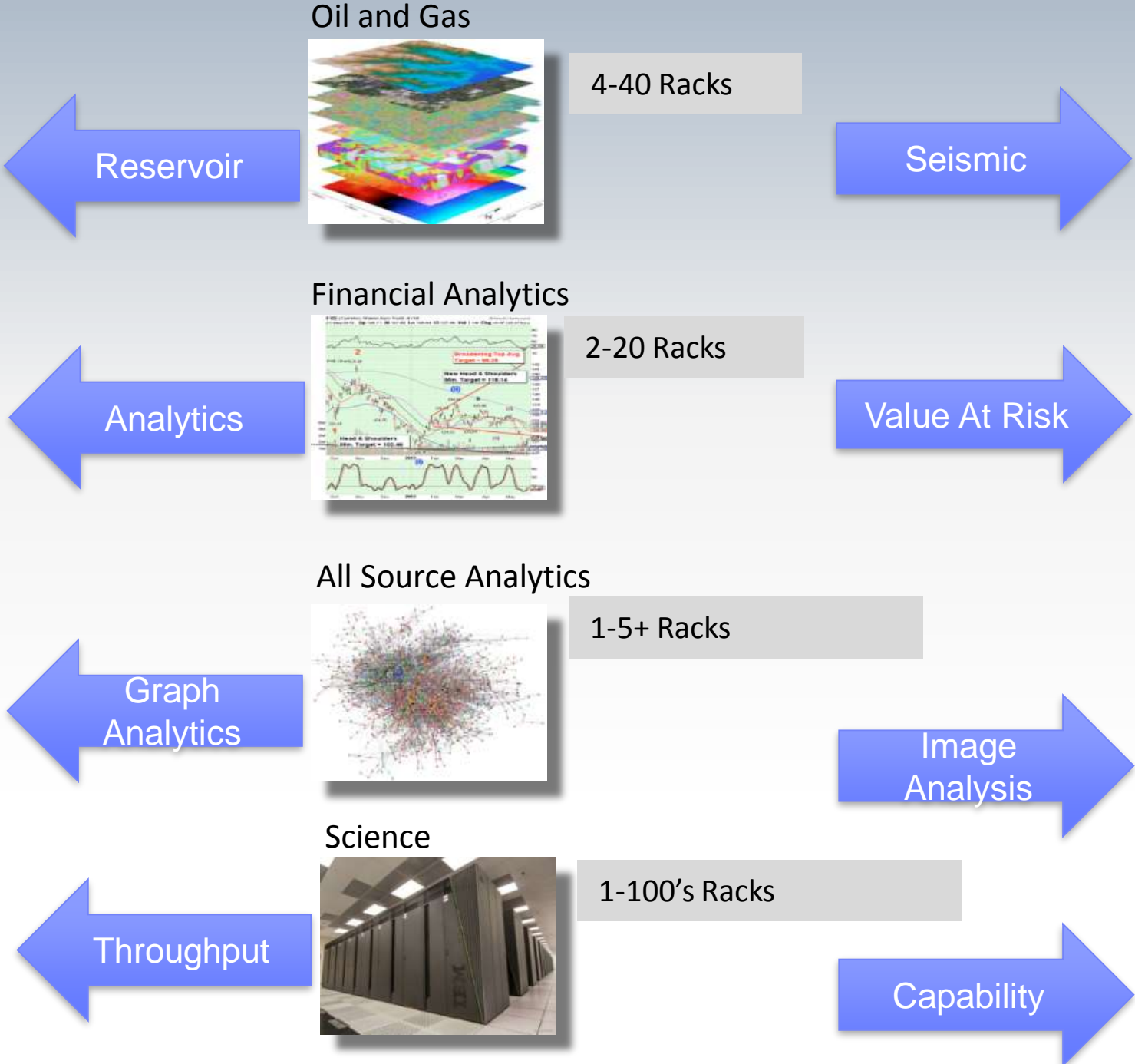
Different Solutions for Different Types of Workflows



Data Centric Workflows: Mixed Compute Capabilities Required

Analytics Capability:

- Complex code
- Data Dependent Code Paths / Computation
- Lots of indirection / pointer chasing
- Often Memory System Latency Dependent
- C++ templated codes
- Limited opportunity for vectorization
- Limited scalability
- Limited threading opportunity

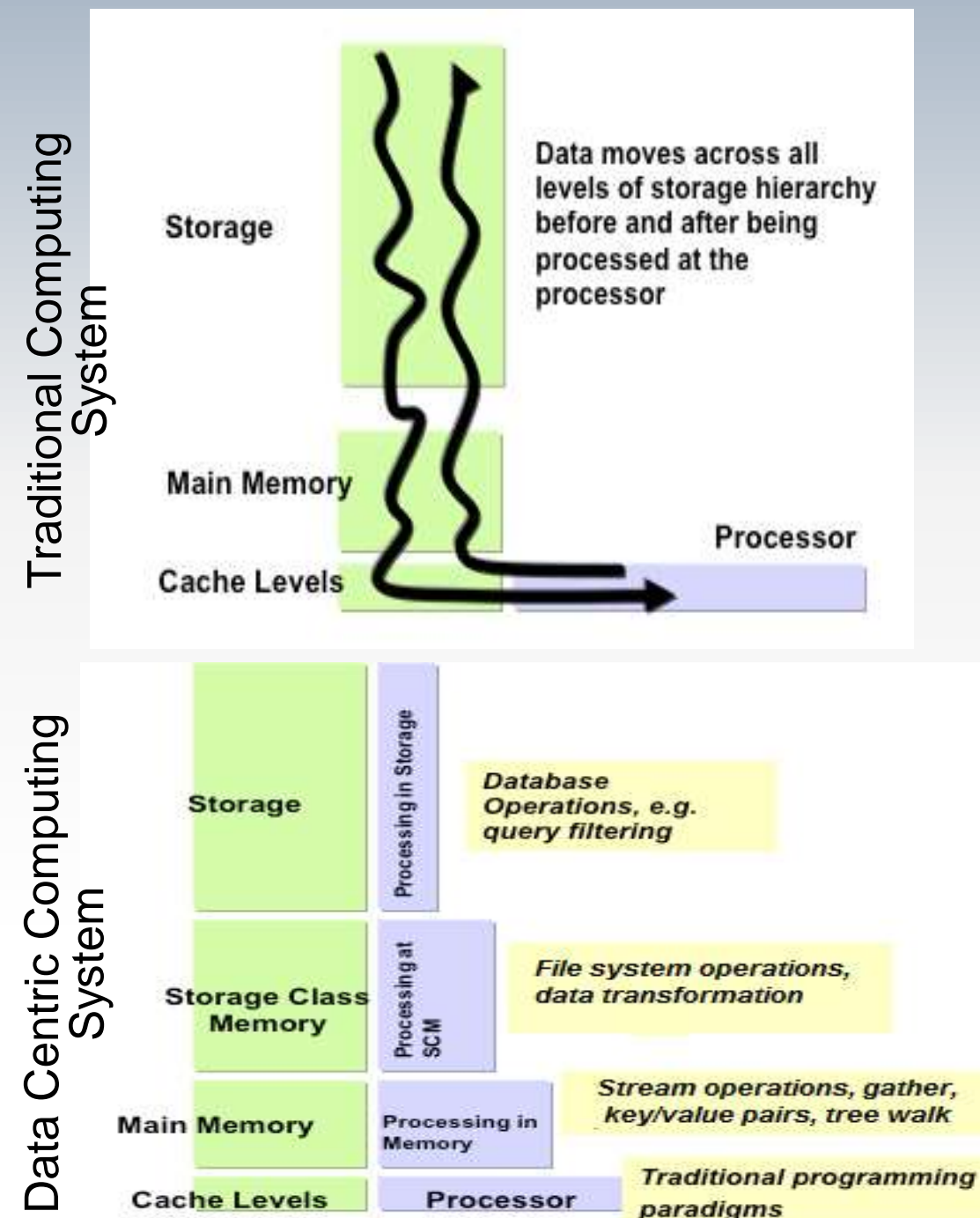


Massively Parallel Compute Capability:

- Simple kernels,
- Ops dominated (e.g. DGEMM, Linpack)
- Simple data access patterns.
- Can be preplanned for high performance.

Comparing Compute Centric to Data Centric

- Systems and Solutions must become more data centric and data aware
 - Data movement minimized
 - Within the system
 - Within/across the end to end solution
 - Compute enabled at all levels
 - Workloads/workflow driven system and solution design choices
 - Modular, composable solution architectures
 - Enhanced resource agility and sharing
- Focus must shift from algorithms to workflows
 - New end to end efficiencies and optimizations
 - Based on a data aware understanding of the full scope of resource requirements
 - Storage, Networking, Compute, Applications, Resource management, Data Centers
- Uncertainty of the future puts a premium on flexibility and innovation



IBM Data Centric Design Principles

Massive data requirements drive a composable architecture for big data, complex analytics, modeling and simulation. The DCS architecture will appeal to segments experiencing an explosion of data and the associated computational demands

Principle 1: [Minimize data motion](#)

- Data motion is expensive
- Allow workloads to run where they run best

Principle 2: [Enable compute in all levels of the systems hierarchy](#)

- HW & SW innovations to support / enable compute in data

Principle 3: [Modularity](#)

- Balanced, composable architecture for Big Data analytics, modeling and simulation

Principle 4: [Application-driven design](#)

- Use real workloads/workflows to drive design points

Principle 5: [Leverage OpenPOWER](#) to accelerate innovation and broaden diversity for clients

IBM OpenPOWER-based HPC Roadmap

Mellanox Interconnect Technology

Connect-IB
FDR Infiniband
PCIe Gen3

ConnectX-4
EDR Infiniband
CAPI over PCIe Gen3

ConnectX-5
Next-Gen Infiniband
Enhanced CAPI over PCIe Gen4

NVIDIA GPUs

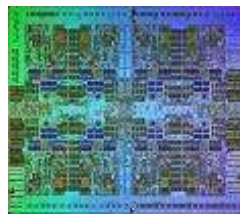
Kepler
PCIe Gen3

Pascal
NVLink

Volta
Enhanced NVLink

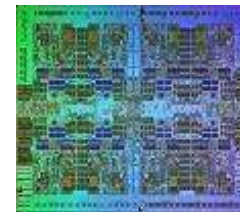
IBM CPUs

POWER8



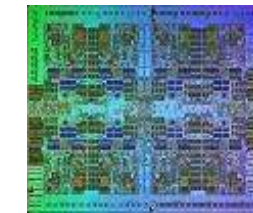
OpenPower
CAPI Interface

POWER8+



NVLink

POWER9



Enhanced
NVLink

Heterogeneity is Key

2015



2016



2017



IBM Nodes

OpenPOWER, a catalyst for Open Innovation

Market Shifts

- Moore's law no longer satisfies performance gain
- Growing workload demands
- Numerous IT consumption models
- Mature Open software ecosystem



Open Development

open software, open hardware



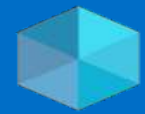
Collaboration of thought leaders

simultaneous innovation, multiple disciplines



Performance of POWER architecture

amplified capability



OpenPOWER™

New Open Innovation

- Rich software ecosystem
- Spectrum of power servers
- Multiple hardware options
- Derivative POWER chips

Feeds back ... resulting in client choice

OpenPOWER is an open development community, using the POWER Architecture to serve the evolving needs of customers.





Implementation, HPC & Research

Software



System Integration



I/O, Storage & Acceleration



Boards & Systems



Chips & SoCs



US & UK Research Centers Select OpenPOWER-based Supercomputers

IBM, Mellanox, and NVIDIA awarded \$325M U.S. Department of Energy's CORAL Supercomputers

CORAL: Leadership Class Supercomputers

5x – 10x HIGHER APP PERF THAN CURRENT SYSTEMS



IBM & UK's STFC in £313M Partnership for Big Data & Cognitive Computing Research



Hybrid CPU/GPU architecture



- At least 5X Titan / Sequoia Application Performance
- Approximately 3,400 nodes, each with:
 - Multiple IBM POWER9™ CPUs and multiple NVIDIA Tesla® GPUs using the NVIDIA Volta architecture
 - CPUs and GPUs completely connected with high speed NVLink
 - Large coherent memory: over 512 GB (HBM + DDR4)
 - All memory directly addressable from the CPUs and GPUs
- Over 40 TF peak performance per node
- Dual-rail Mellanox® EDR-IB full, non-blocking fat-tree interconnect
- IBM Elastic Storage (GPFS™) - 1TB/s I/O and 120 PB disk capacity.



Programming Approaches

- Accelerator Approach:
 - Required when not coherent
 - Each processor computes in its own private address space
 - Data objects are homed in CPU Memory, are copied to GPU memory for GPU execution, GPU engines act only on data in GPU memory.

- Compute in Shared Address Space:
 - New option, now that CPU / GPU memories are coherent
 - Data objects can be in any physical memory domain
 - Processors (either CPU or GPU) can use data in place.
 - No copies required

- Note:
 - Will still have to manage NUMA

Compute and Memory View – Emerging Approach

GPU 1 Compute
GPU 1 Memory

GPU 2 Compute
GPU 2 Memory

CPU 1 Compute
CPU 1 Memory

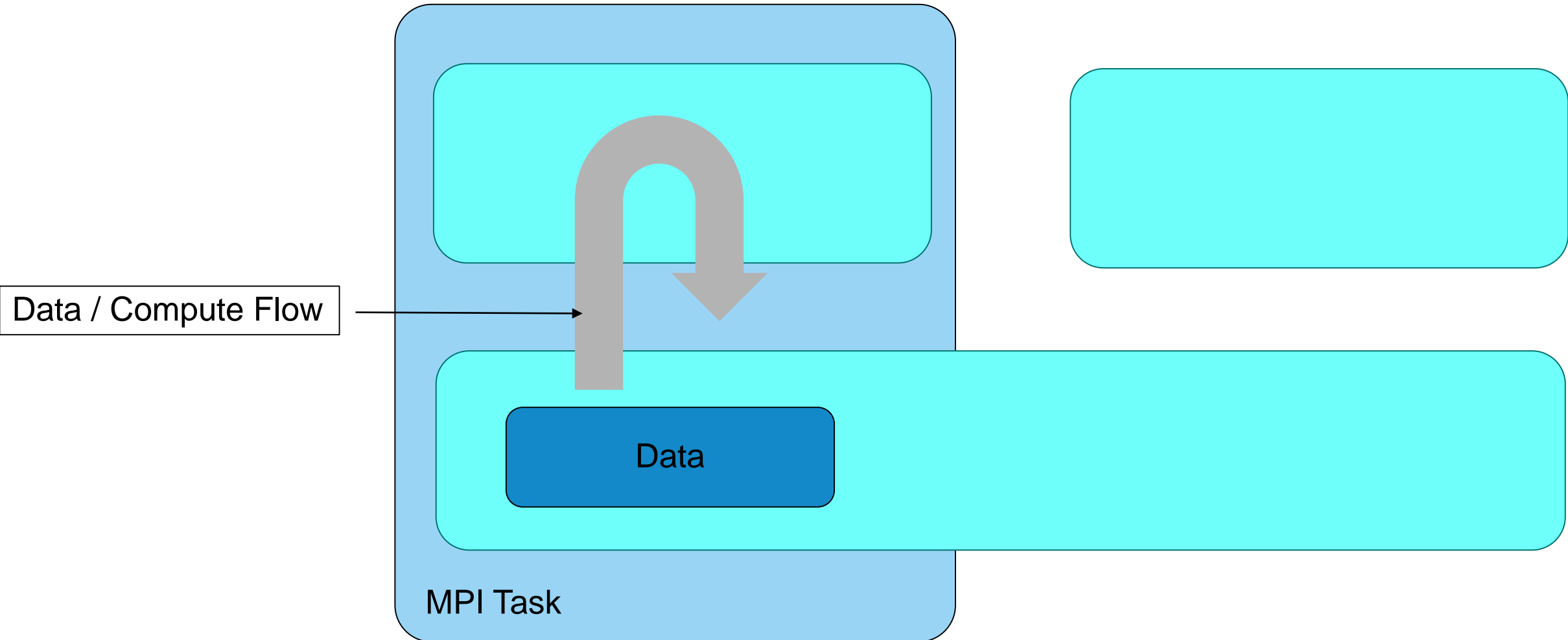
Multiple Compute Engines

- Consider all engines as equal peers

Multiple Memories

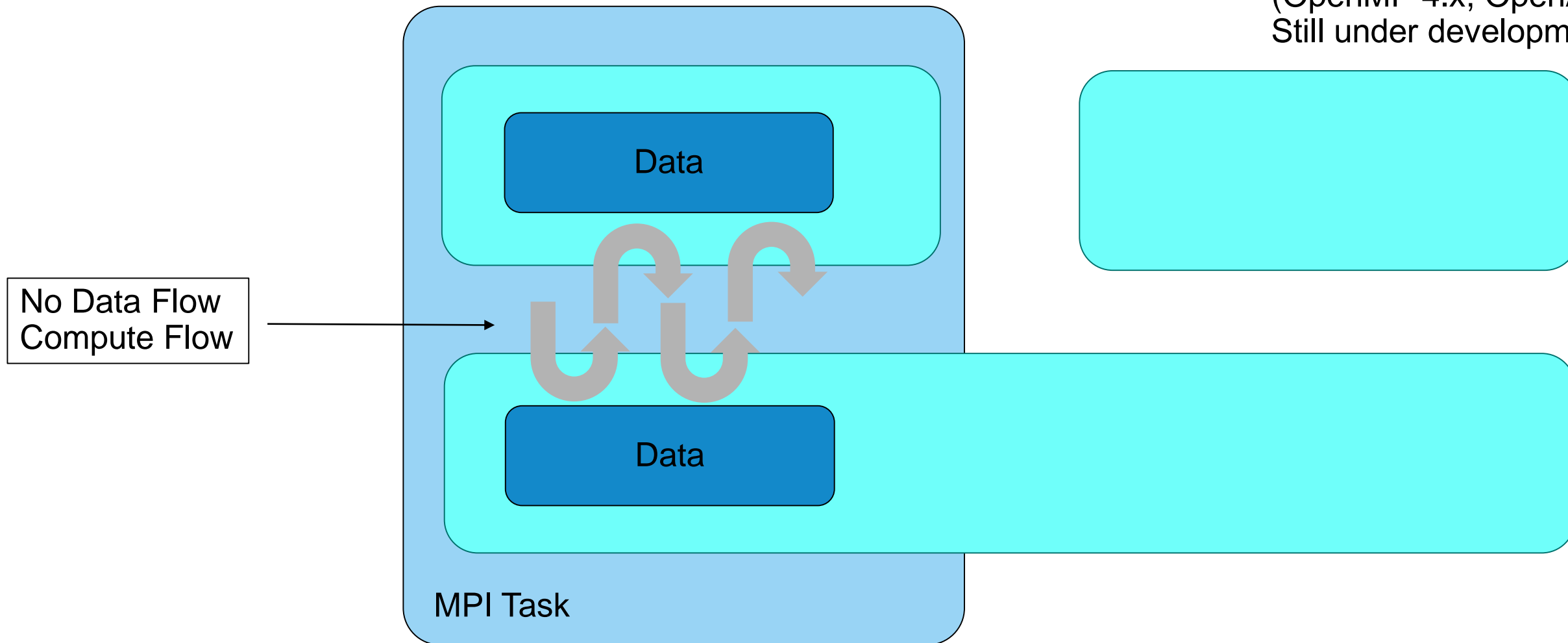
- Consider all memories as equal peers

Compute and Memory View – Traditional Acceleration

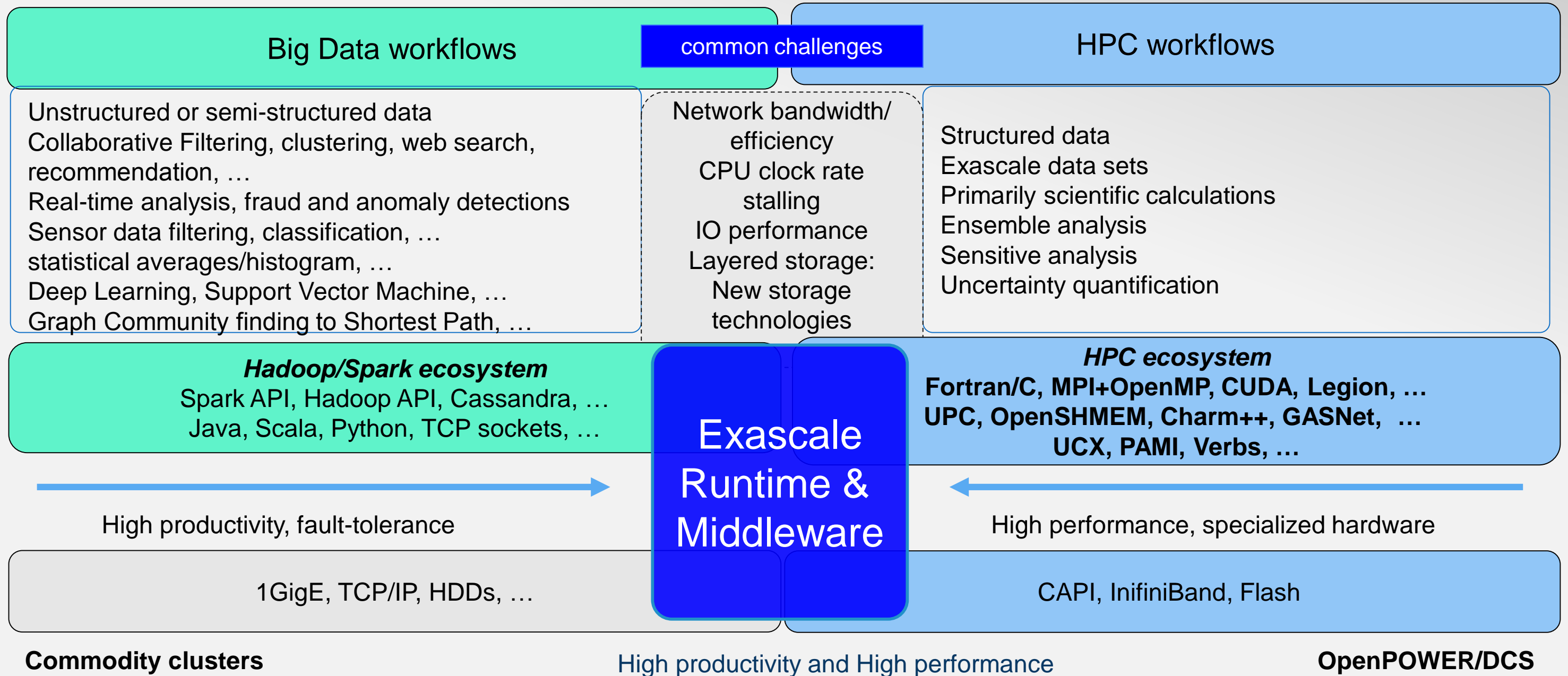


Peer Processing

Data can be placed at Allocation, or
 can migrate under run time control
 (e.g. UVM)
 Thread-like programming model
 (OpenMP 4.x, OpenAcc, CUDA ...
 Still under development ...



Future DCS Programming Model



DCS OpenPOWER Software Contribution Examples

- Linux – NUMA support for hardware coherent GPU memory
- Provisioning – xCAT
- Burst Buffer – Support for fast shared-file checkpointing
- CSM – Cluster System Management
- Compilers – LLVM
- Tools – Ensure tools have appropriate APIs

Summary – IBM Data Centric Systems & OpenPOWER

- Experience/Observation: Extraction of meaningful insights from Big Data and enabling real-time, predictive decision making requires similar computation techniques that have been characteristic of Technical Computing
 - Convergence in many future workflow requirements including Big Data-driven analytics, modeling, visualization, and simulation
 - Will require optimized full-system design and integration
- IBM Approach: Data Centric innovation in multiple areas, in open ecosystems with workload-driven co-design
 - System architecture and design with modular building blocks
 - Hardware technologies
 - Integration of heterogeneous compute elements
 - Software enablement