Paper Review: Network Support for Resource Disaggregation in Next-Generation Datacenters

Tanjila Ahmed

UC DAVIS
UNIVERSITY OF CALIFORNIA

Outline

- Big Data Challenges for Datacenter Network
- Evolution of Datacenter Architecture
- Server Centric Datacenter
- Resource Centric Datacenter
- Resource Requirement
- Trends
- Proposed Disaggregated Datacenters
- Assumptions
- Latency and Bandwidth Requirement
- Making Memory Traffic Manageable
- Experiment
- Findings

UCDAVIS

References

[1] S. Han, N. Egi, A. Panda, S. Ratnasamy, G.Shi, and S. Shenker, "Network Support for Resource Disaggregation in Next-Generation Datacenters", Hotnets '13, November 21–22, 2013.

[2] Huawei technical white paper, High Throughput Computing Data Center Architecture (2014) [Available Online] http://www.huawei.com/ilink/en/download/HW_349607&usg=AFQjCNE0mKD71dxJeRf1cJSkNaJbpNgnw&cad=rja.

UCDAVIS

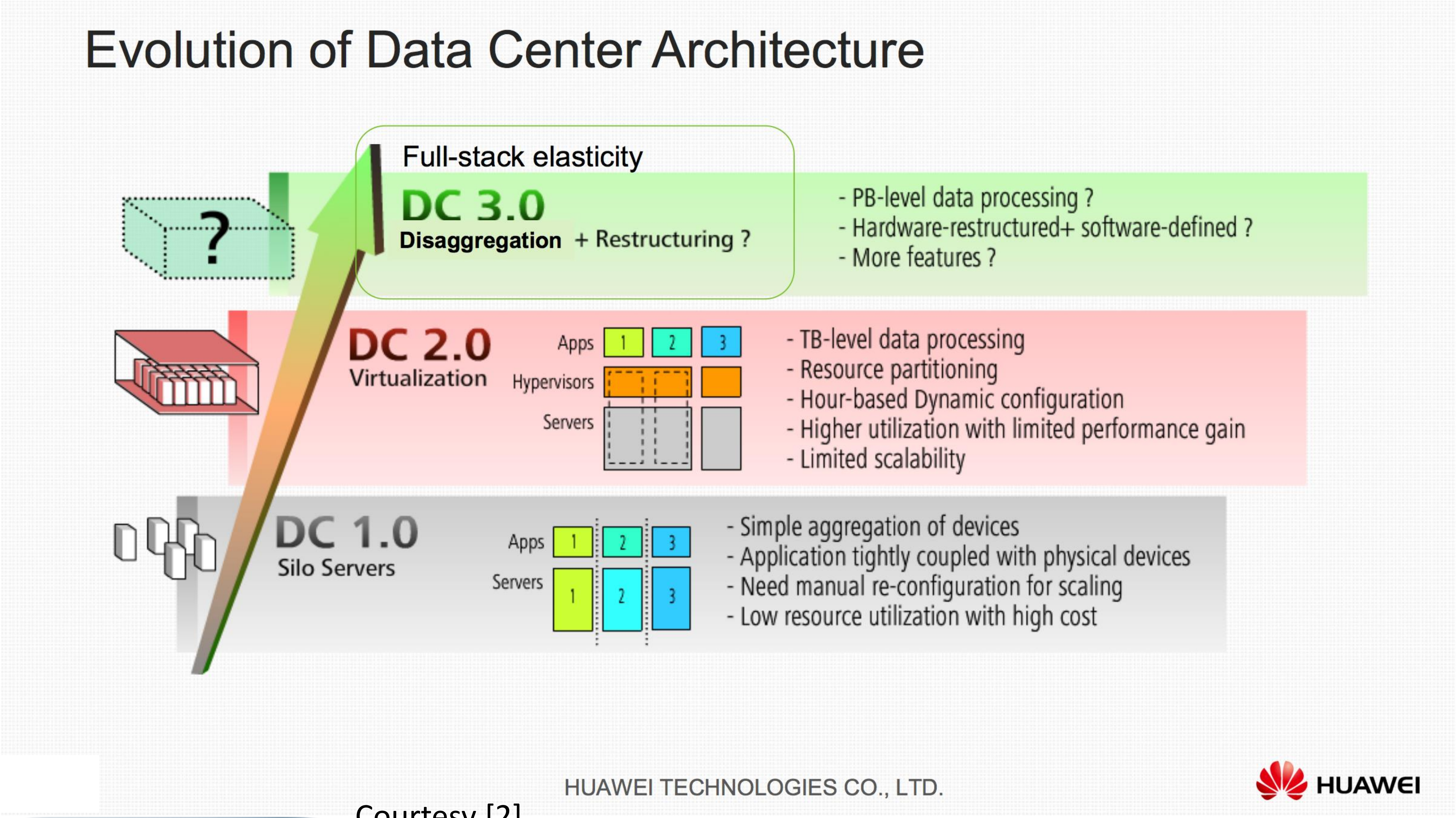
Big Data Challenges to Data Centers

Limitations of Current DC

Throughput	Resource Utilization	Management	Scalability	Energy Efficiency
• Data processing capability • I/O bottleneck	• Typically Utilization<30% • Virtualization with high overhead	• Limited flexibility for deployment and configuration • Complex operations	• High speed copper interconnect • DC-level large-scaled interconnect	• Lower power efficiency
• New medium • New architecture • New access Mechanism	• Resource disaggregation • On-demand and flexible resource allocation	• Intelligent Management • Self-healing • Self-configuration • Software-defined	• Optics based interconnect	• New architecture for energy efficient computing

Strategies

Courtesy [2]

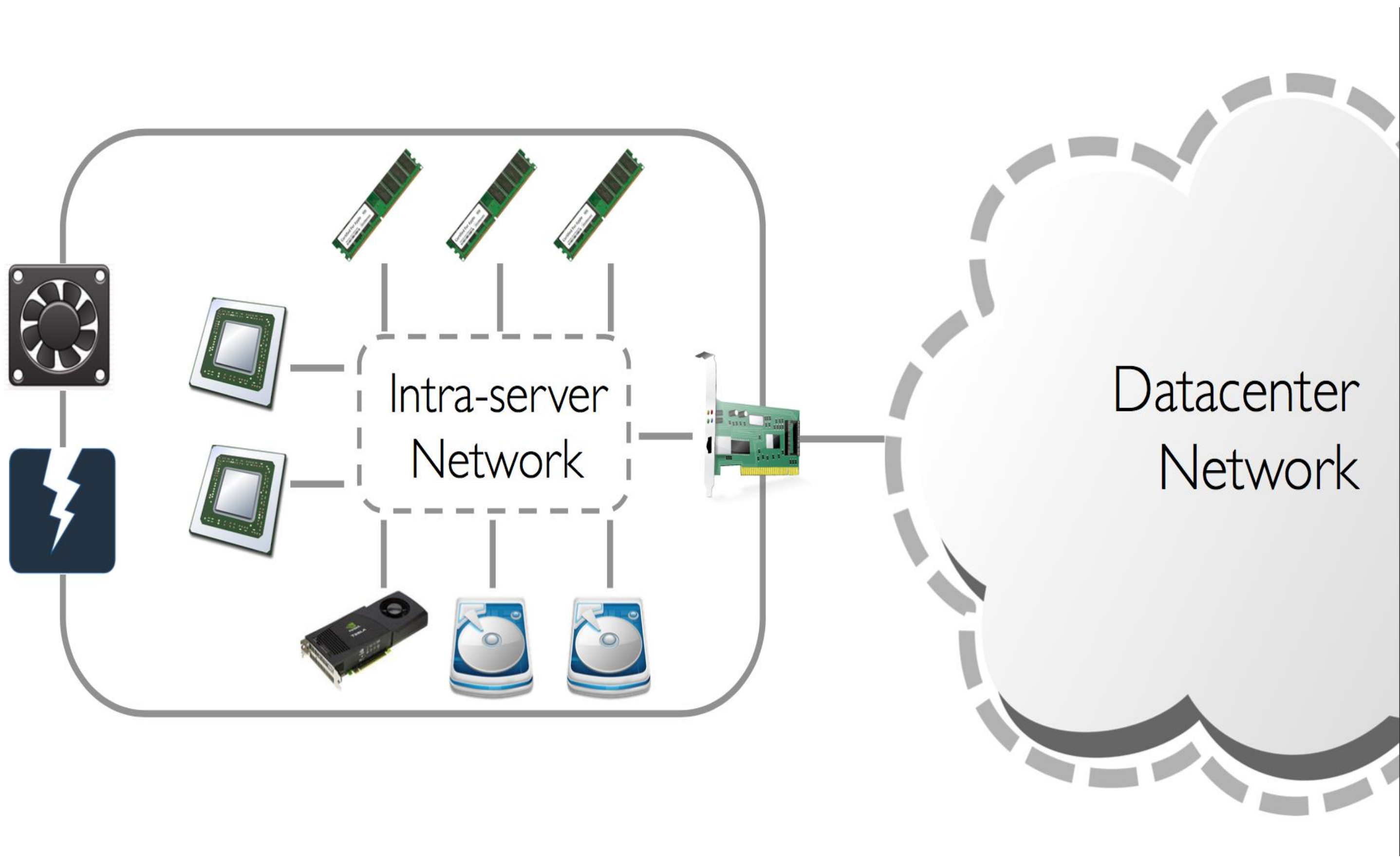


Courtesy [2]

UCDAVIS

Server Centric Datacenter

- Each server aggregates a fixed amount of computing, memory, storage, and communication resources.



Resource Centric Datacenter

- Aggregation of resources is logical(allocated by a software scheduler) rather than physical(dictated by hardware)
- Physically decoupling resources
- Allows each technology to evolve independently & provides fine-grained control over selection, provision, & upgrade individual resources.

Resource Requirement

- Figure 1 plots the ratio of disk-to-CPU and memory-to-CPU consumption for tasks in Google’s datacenter
- It shows that the resource requirements of tasks vary greatly.

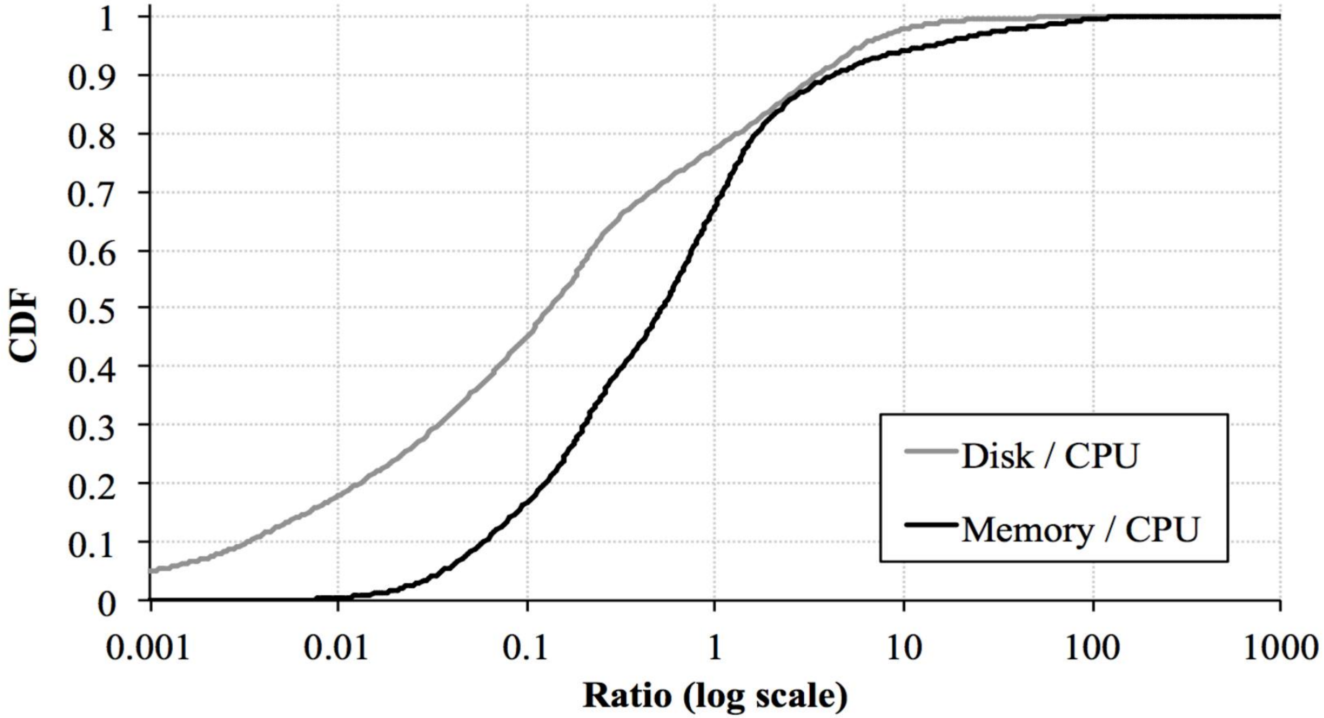


Figure 1: Distribution of disk/memory capacity demand to CPU usage ratio for tasks in Google’s datacenter.

UCDAVIS

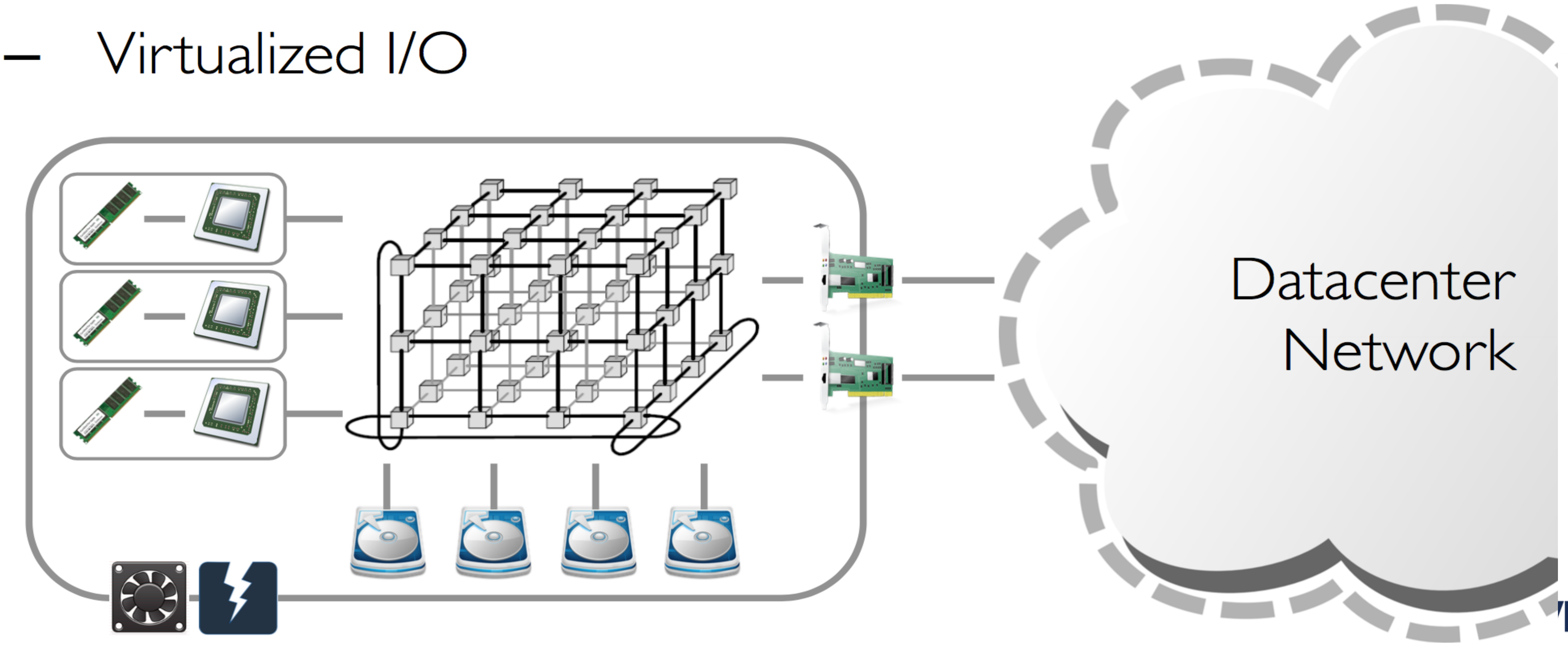
Trends: Disaggregation

1. HP MoonShot
   – Shared cooling/casing/power/mgmt for server blades

UCDAVIS

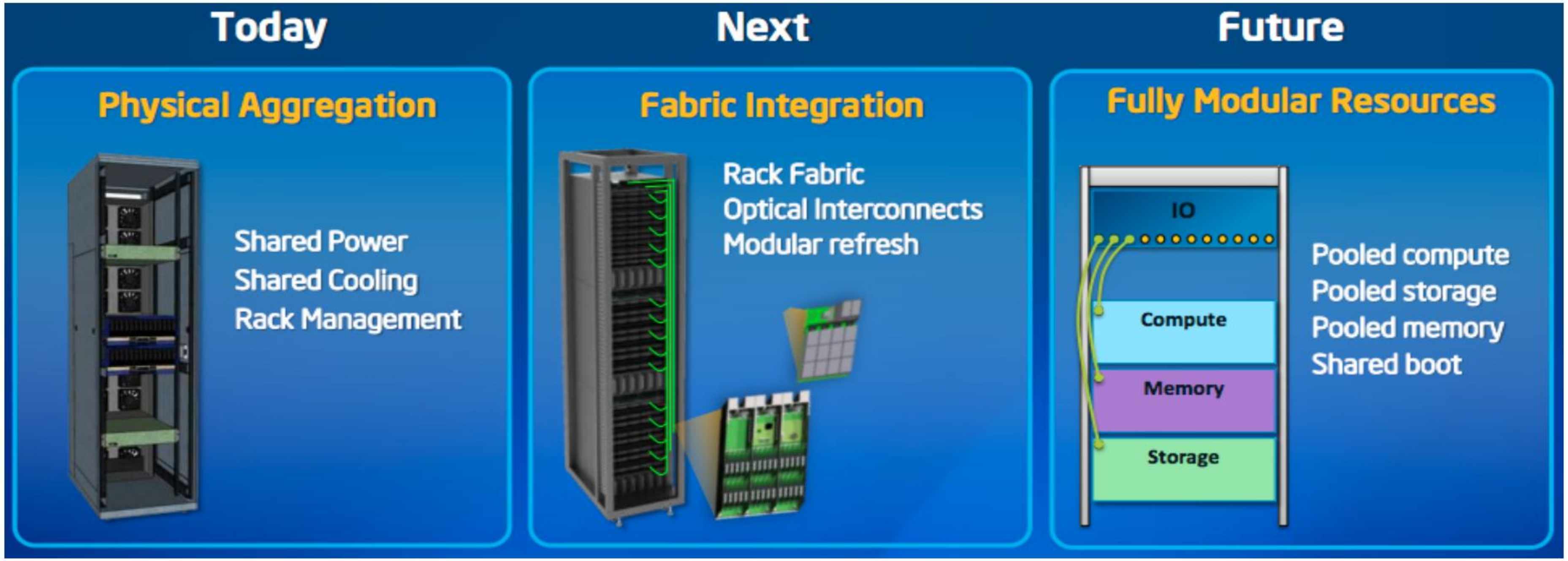
Trends: Disaggregation

1. HP MoonShot
2. AMD SeaMicro
   – Virtualized I/O



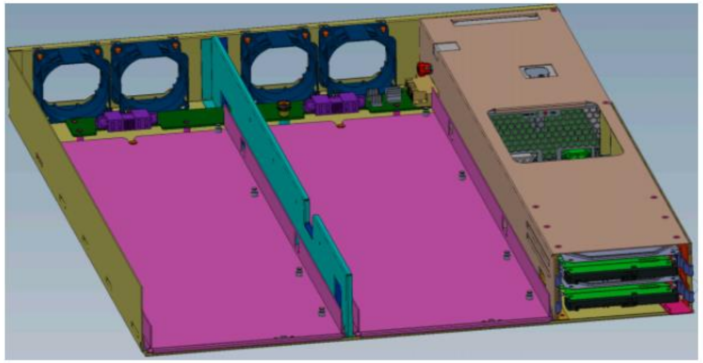
Trends: Disaggregation

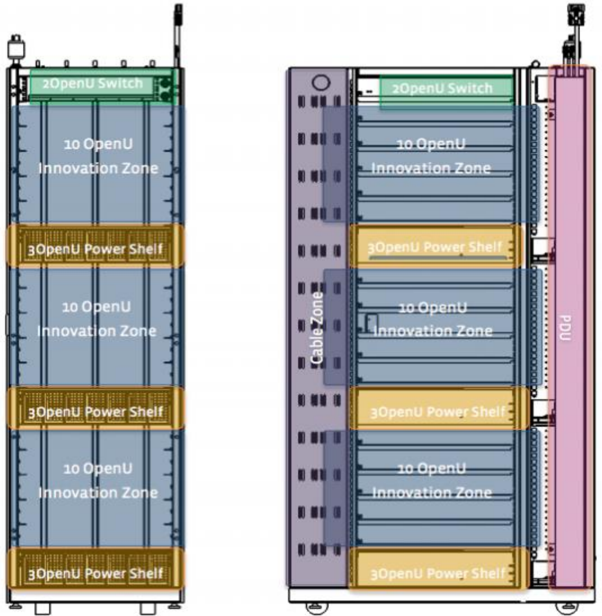
1. HP MoonShot
2. AMD SeaMicro
3. Intel Rack Scale Architecture



The Trends: Disaggregation

1. HP MoonShot
2. AMD SeaMicro
3. Intel Rack Scale Architecture
4. Open Compute Project





Trends: Disaggregation

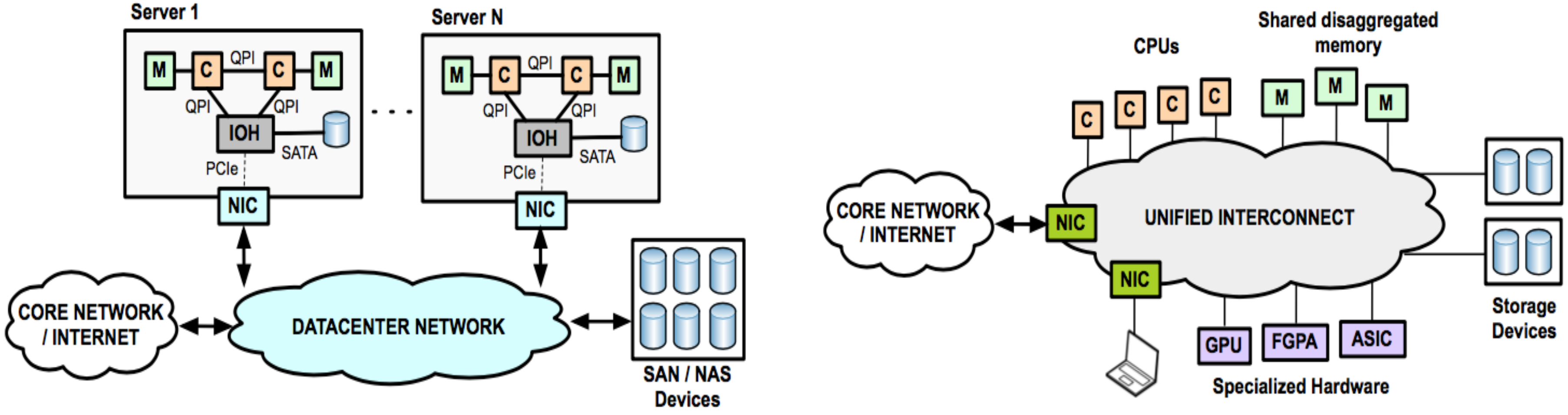
5. Facebook Open Switching System (FBOSS): distributing the switches functionalities across the whole network.

6. High Throughput Computing Data Center (HTC-DC) Architecture from Huawei : focuses on a disaggregated DC architecture where blades are interconnected through a high bandwidth optical network fabric.

UCDAVIS

Proposed Disaggregated Datacenters

Resource as a standalone blade



(a) Current datacenter

(b) Disaggregated datacenter

Figure 2: Architectural differences between server-centric and resource-centric datacenters

Proposed Disaggregated Datacenters

- HW Requires Minimal Modification

  The internals don't need to change.

  All we need is embedded network controller.

  – They already have: QPI, HT, PCIe, SATA,...

  – Can be very cheap

  E.g., a whole graphics card w/ 128Gbps for only $50

- Existing SW infrastructure heavily relies on the concept of "server"

  -We don't want to rewrite it from scratch.

  -No modification for App/OS

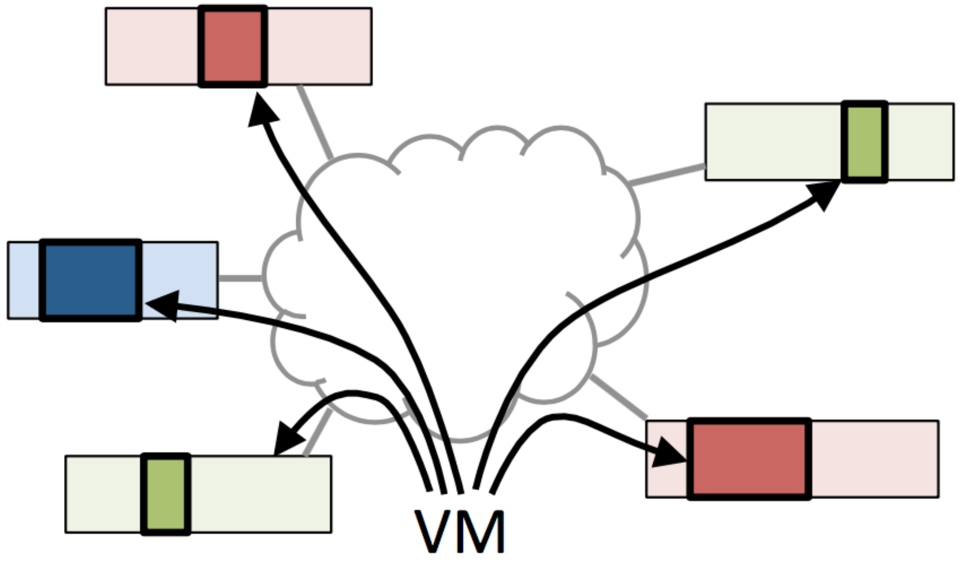
  -Minor changes in VMM.

  -Much higher utilization!

UCDAVIS

Proposed Disaggregated Datacenters

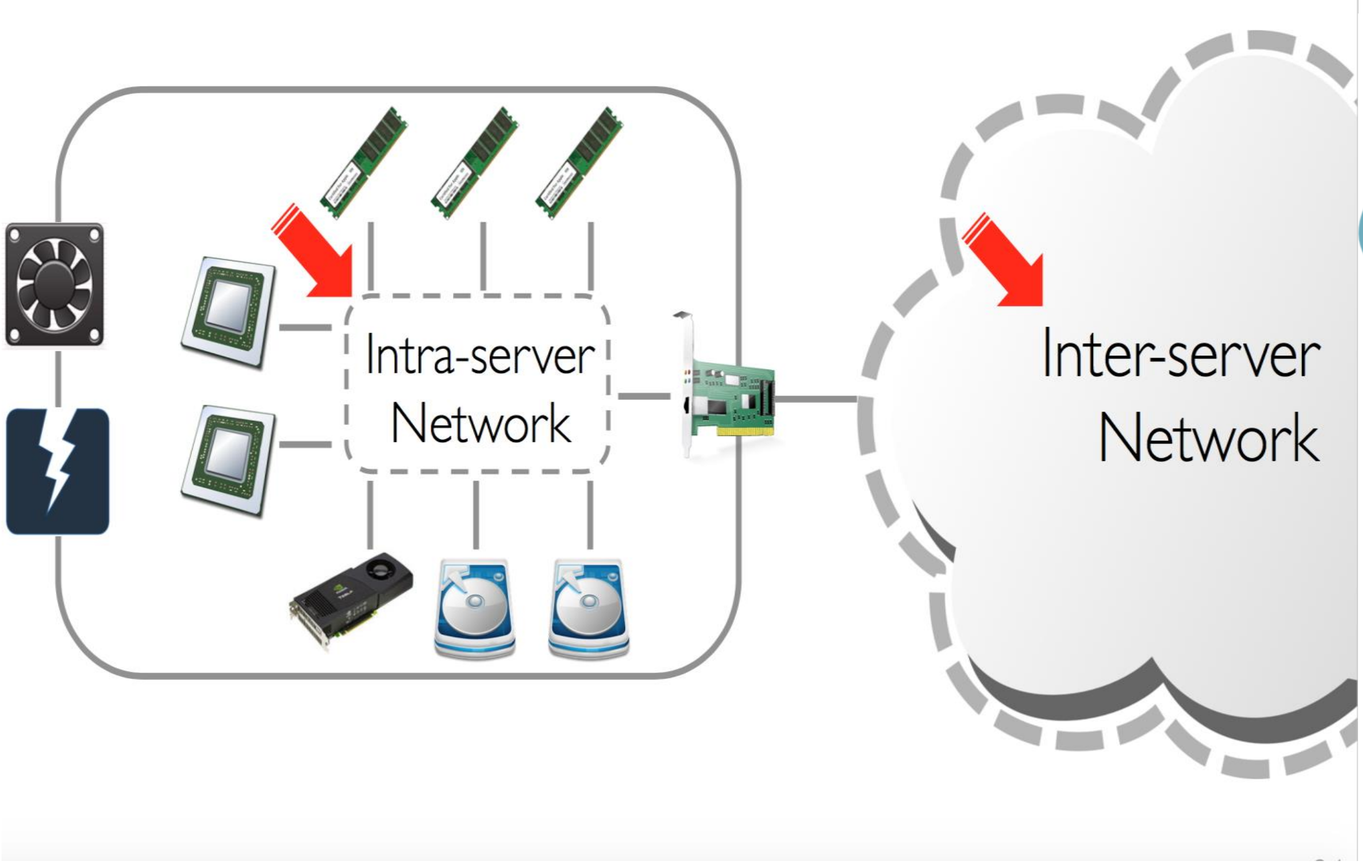
- Elastic VMs Achieve High Utilization!
  1. No “server boundary”
  2. Statistical multiplexing at a larger scale
  3. Higher utilization!



Disaggregated VM

UCDAVIS

An Unified Network is Plausible



Aren't they two different things?

Not really.

E.g., PCIe and 10GbE

- Serial
- Point-to-point
- Full duplex
- Packet-switched
- Variable packet size
- Supports both message and read/write semantics

No fundamental™ difference!

UCDAVIS

Assumptions

- **VM as a computational Unit**: we assume that computational resources are still utilized by aggregating them to form VMs, while each resource is now physically disaggregated across the datacenter.

- **Local/remote memory**: Since memory access from CPUs must run at very high speed. Each CPU blade retains some amount of local memory that acts as a cache for remote memory. While remote memory may be allocated to any CPUs in the datacenter, local memory is dedicated to its co-located CPU.

UCDAVIS

Assumptions

- **Page-level remote memory access :**

1. CPU blades access remote memory at the page-granularity (4 KB in x86) over the fabric.
2. In addition, page-level access requires little or no modification to the virtual memory subsystem of hypervisor or operating system, and it is completely transparent to user-level applications.
3. Remotely accessed pages are not shared by multiple VMs at a given time, in order to not introduce cache coherence traffic across the network.
4. In paging operation there are two main sources of performance penalty: *i)* software overhead for trap and page eviction and *ii)* page transfer time over the network.

UCDAVIS

Latency and Bandwidth Requirement

Communication type	Latency (ns)	Bandwidth (Gbps)
CPU - CPU	10	200
CPU - Memory	20	300
CPU - 10G NIC	$> 10^3$	10
CPU - Disk (SSD)	$> 10^4$	5
CPU - Disk (HDD)	$> 10^6$	1

Table 1: Typical latency and peak bandwidth requirements within a traditional server. Numbers vary between hardware.

For I/O traffic such as network interfaces& disks, the required latency & bandwidth level is low to consolidate them within unified network.

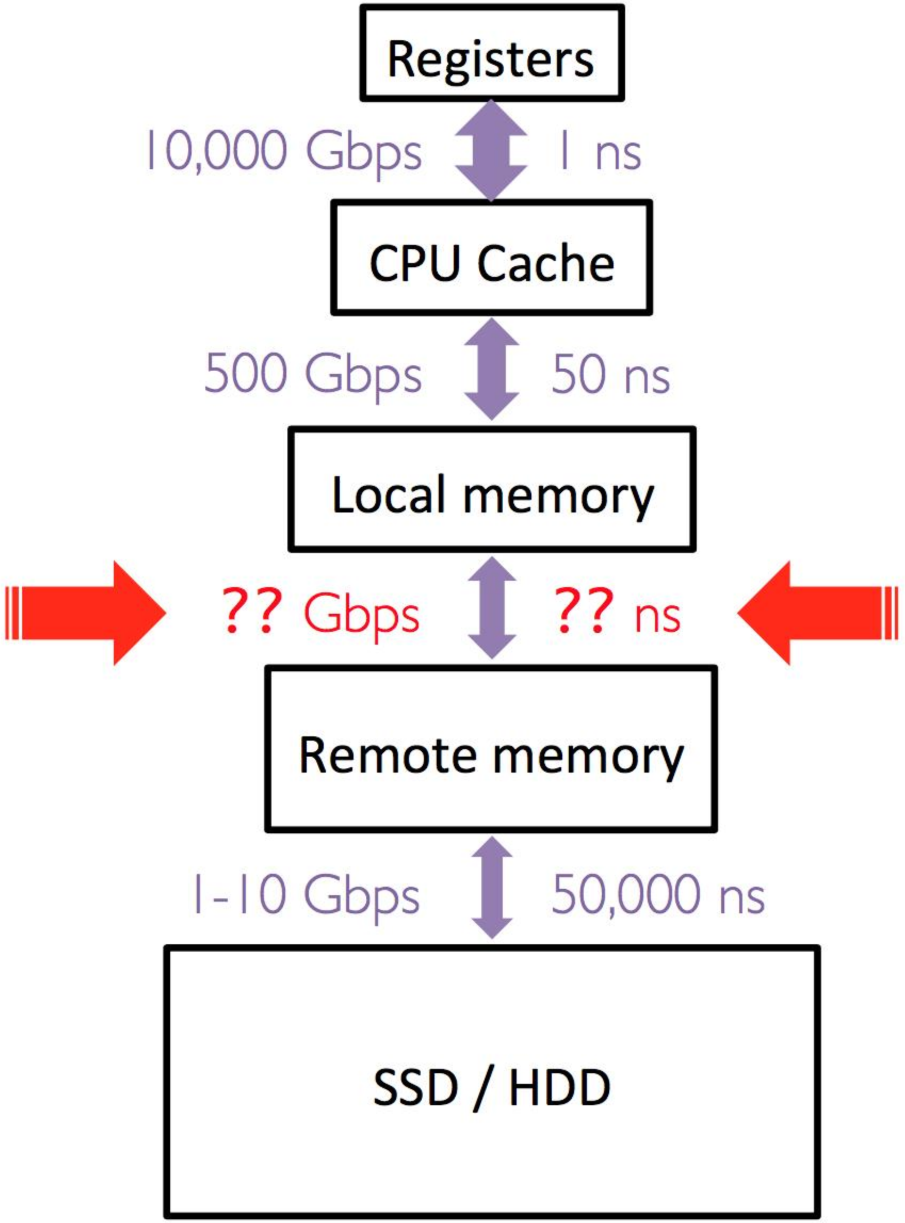
CPU-to-CPU and CPU-to-memory has high bandwidth & extremely low latency requirements.

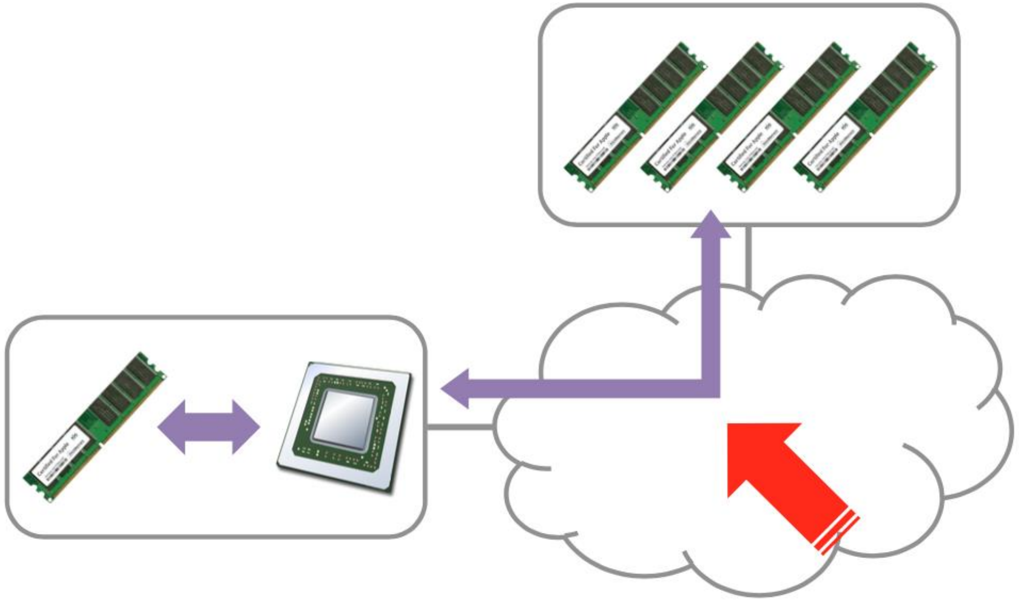
To Avoid those two traffic:

1. Keep each VM from spanning multiple CPU blades, to eliminate CPU-to-CPU traffic.
2. Instead of fully disaggregating memory, we envisage that each CPU has a small amount of private, directly connected local memory.

UCDAVIS

Making Memory Traffic Manageable





A small amount of
local memory as a "cache"

DAVIS

Experiment

Objective: How network latency & bandwidth affect application performance with remote memory access.

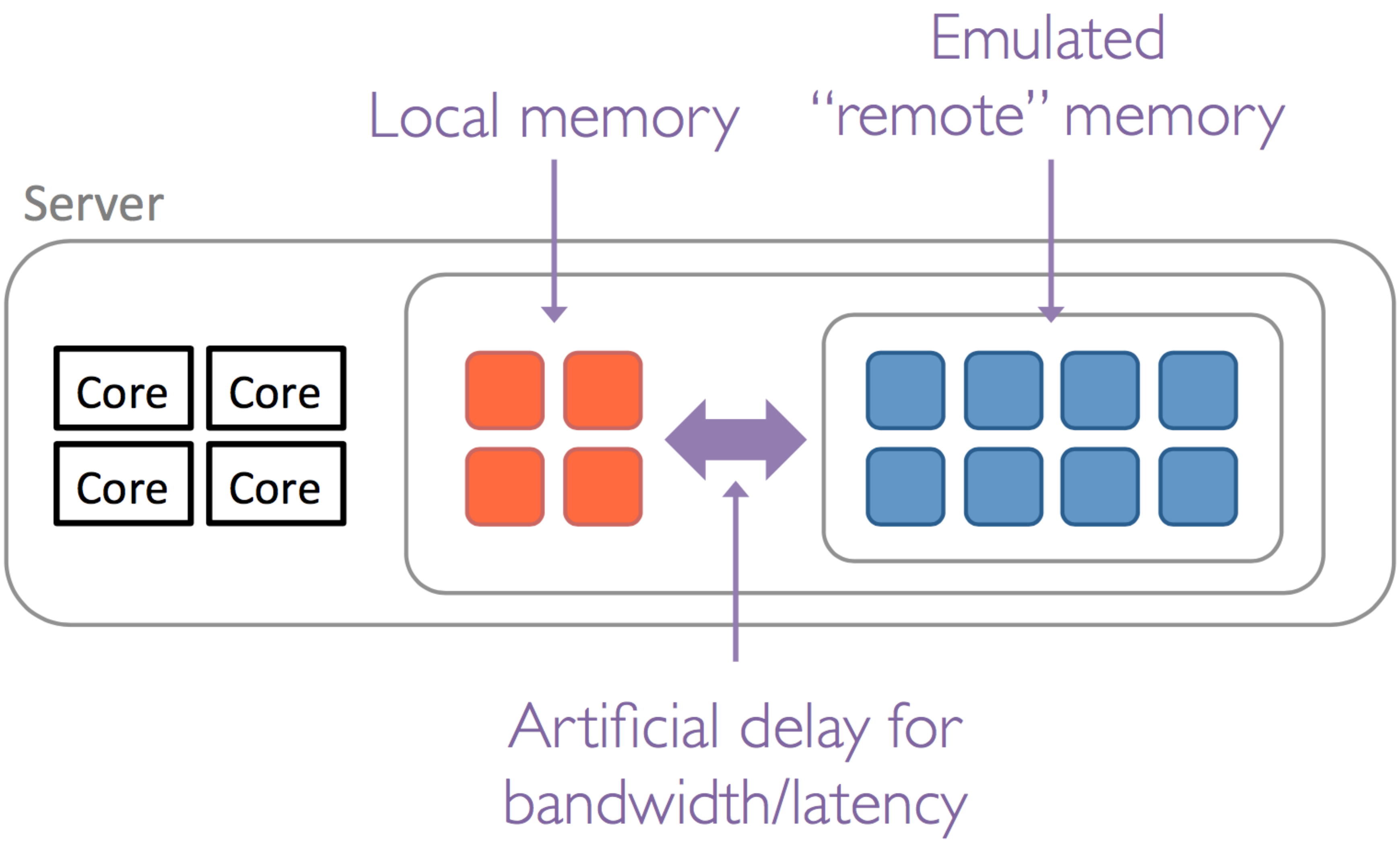
Traffic: GraphLab, a machine learning toolkit; Memcached , an in-memory, key-value store & Pig, a data-analysis platform based on Hadoop.

Method: A remote memory access is implemented using a special swap device (backed by physical memory rather than a disk) & injecting artificial delays to emulate network round-trip latency & bandwidth for each paging operation.

Measurement: Measure relative performance on the basis of throughput or completion time as compared to the zero-delay case. Results do not account for the delay caused by software overhead for page operations.

UCDAVIS

Experiment


Local memory
Emulated
"remote" memory
Server
Core
Core
Core
Core
Artificial delay for
bandwidth/latency

Results

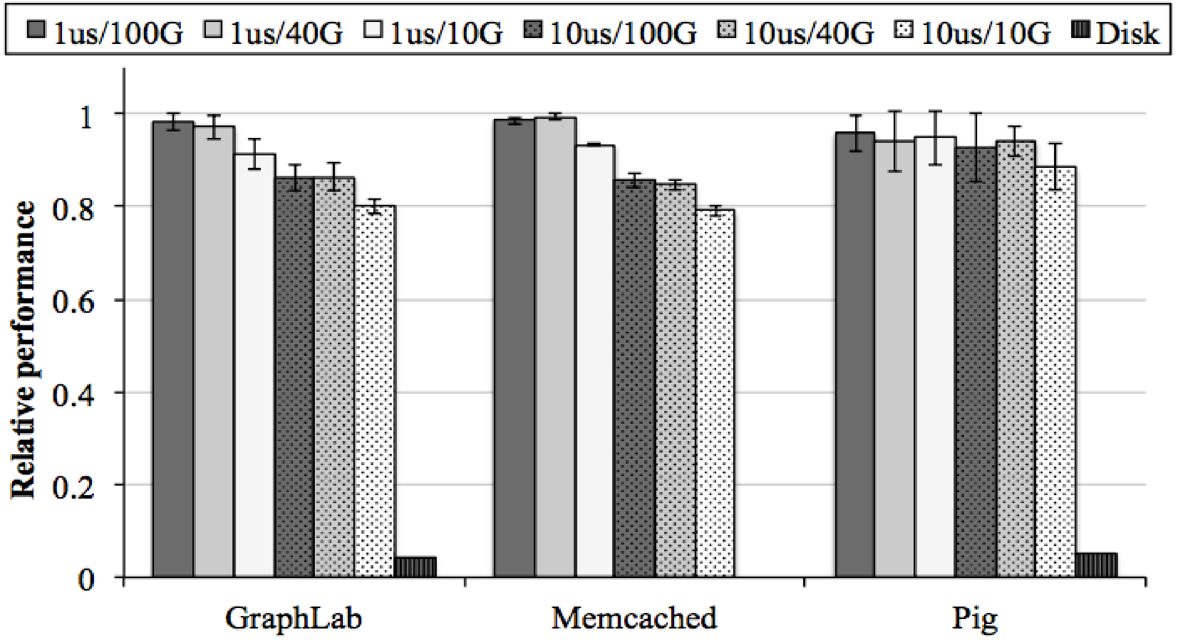
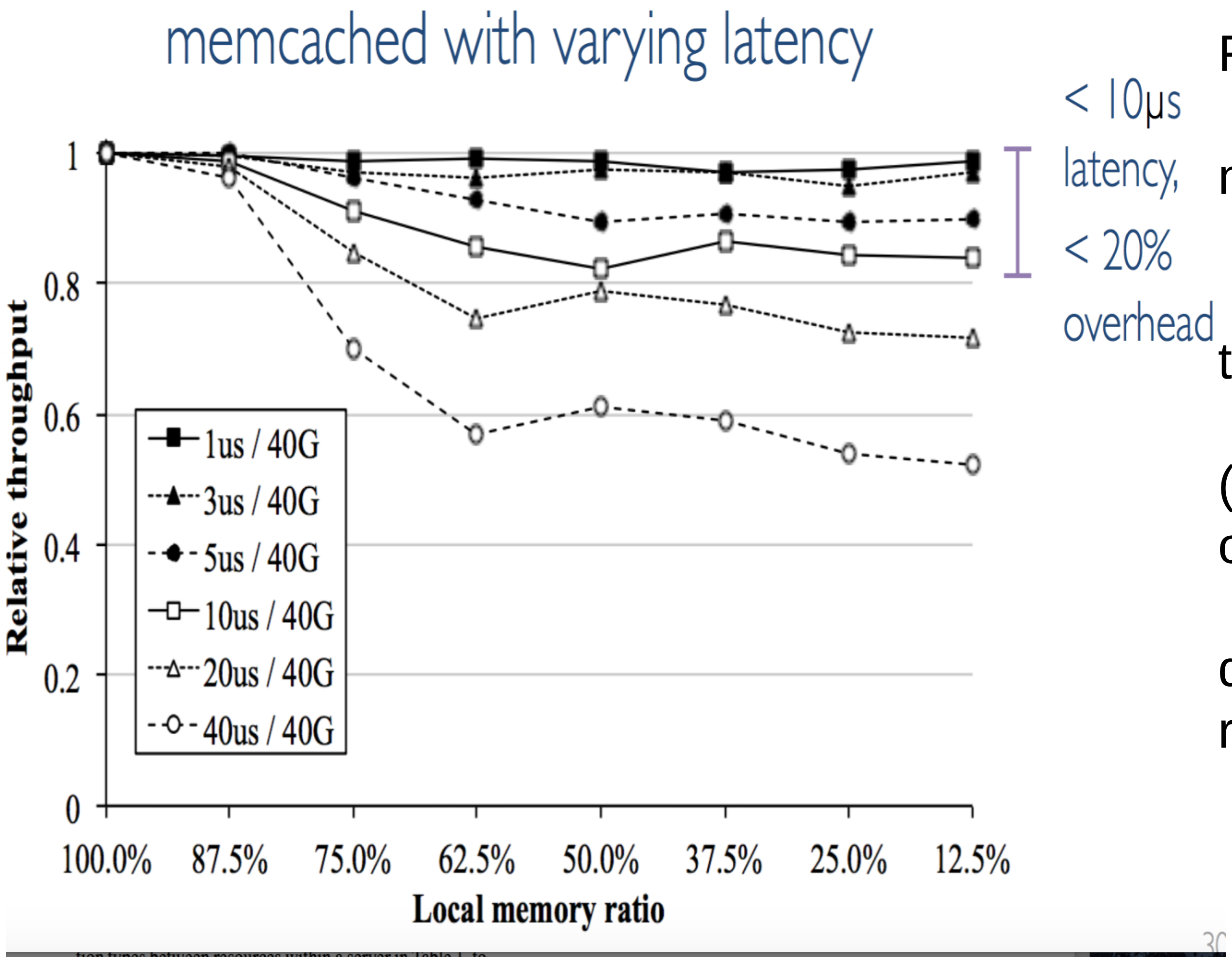


Figure 3: Application-level performance degradation with disaggregated memory, over various network configurations. 75% of the working set size was configured as remote memory. Memcached with disk-based swap performed too slow to get the benchmark result.

1. Use of remote memory can drastically improve application performance when the working set size is bigger than physical memory, as compared to traditional disk-based swap.
2. Second, low latency is more important than high band-width. The 100 Gbps bandwidth did not provide any significant improvement over the 40 Gbps link. In contrast, 10 μs round-trip latency causes noticeable performance degradation, as compared to the 1 μs case.

UCDAVIS

Results



For the experiment, we fixed the band- width at 40 Gbps and varied the amount of local memory from 1 GB to 8 GB, out of the total 8 GB working set size. Figure 4 again confirms that low latency will be crucial in the implementation of resource disaggregation. The low latency (10 μs) cases show fairly constant performance over any local memory ratio, while the performance of high latency (20 μs) cases quickly degrades as we rely more on remote memory.