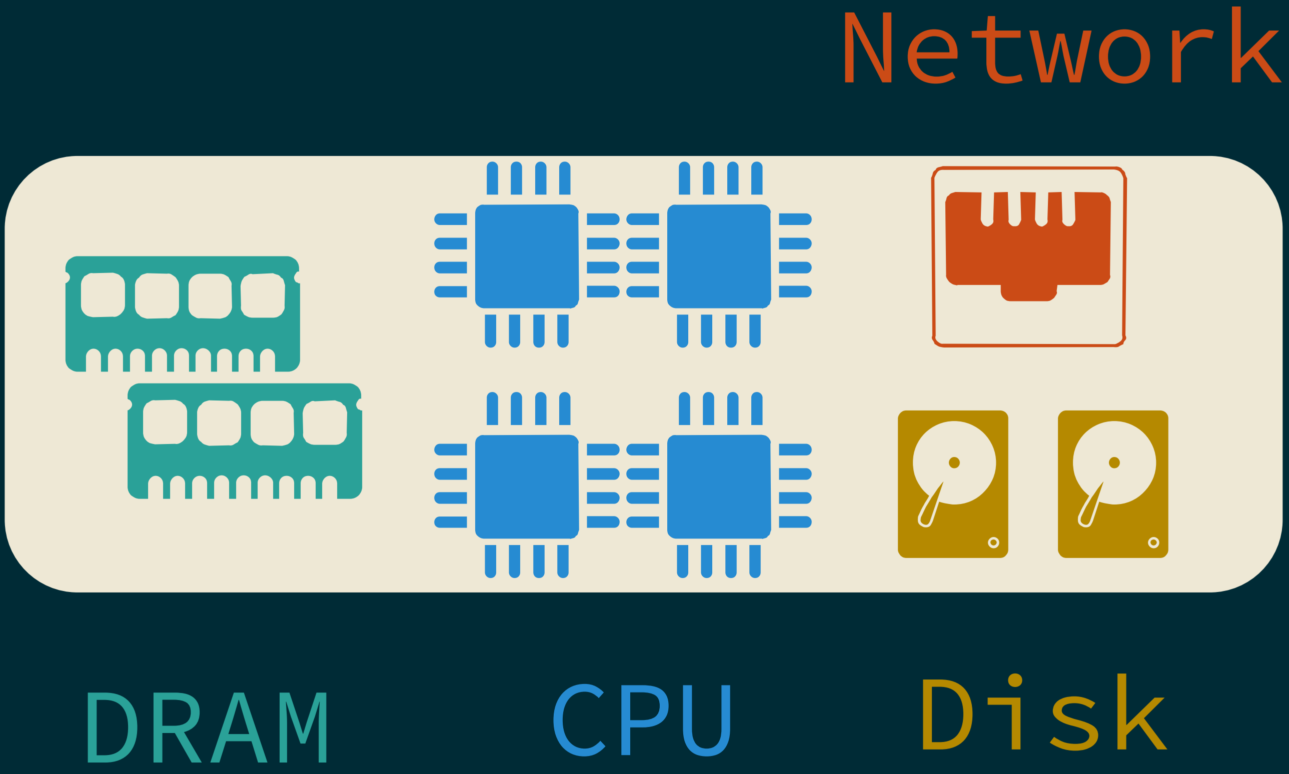
A Peer-to-Peer, Local Area Disaggregated System

Michael Wei, George Porter, Steven Swanson

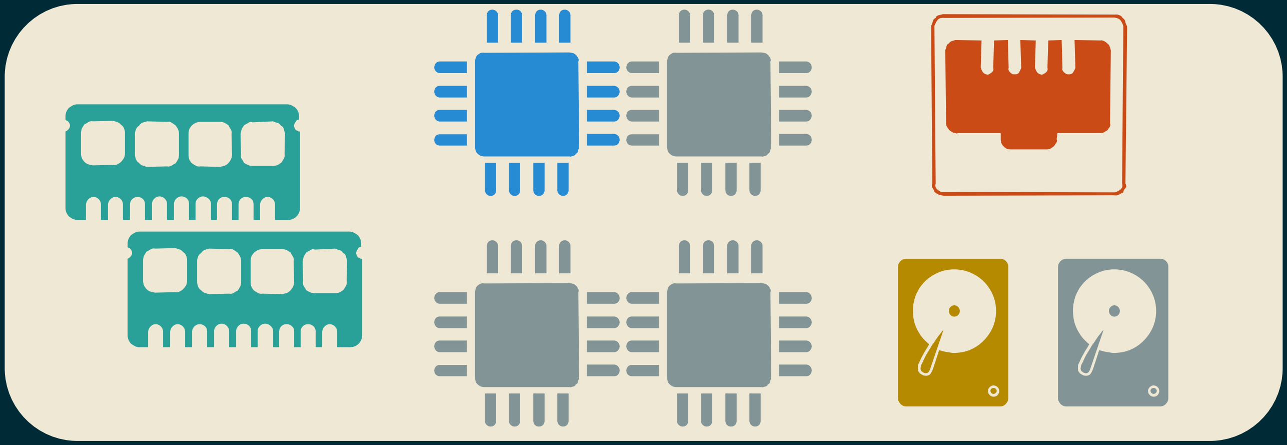
10/22/2014

Disaggregation 〉 Intro



Disaggregation > Intro

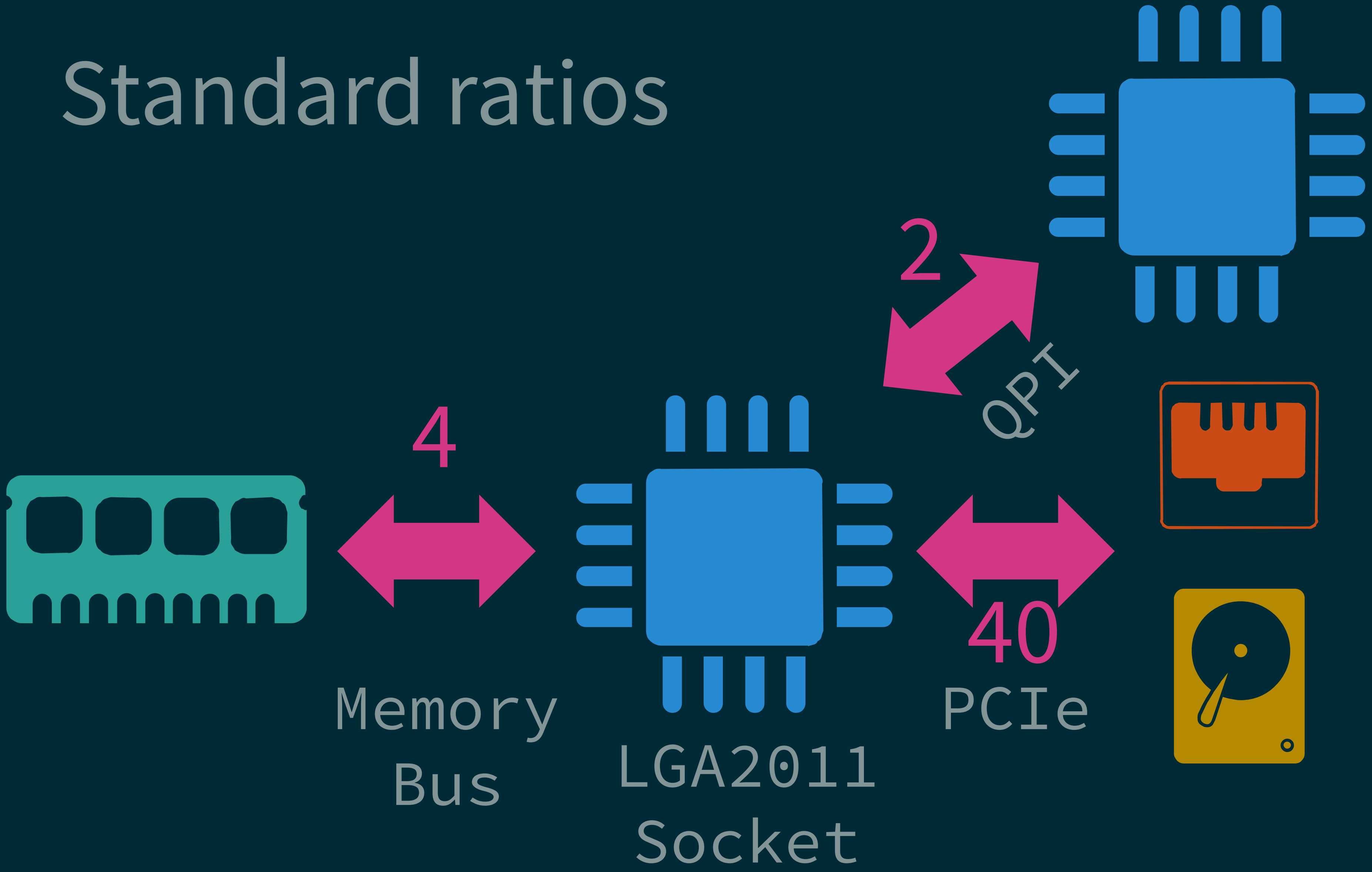
Network



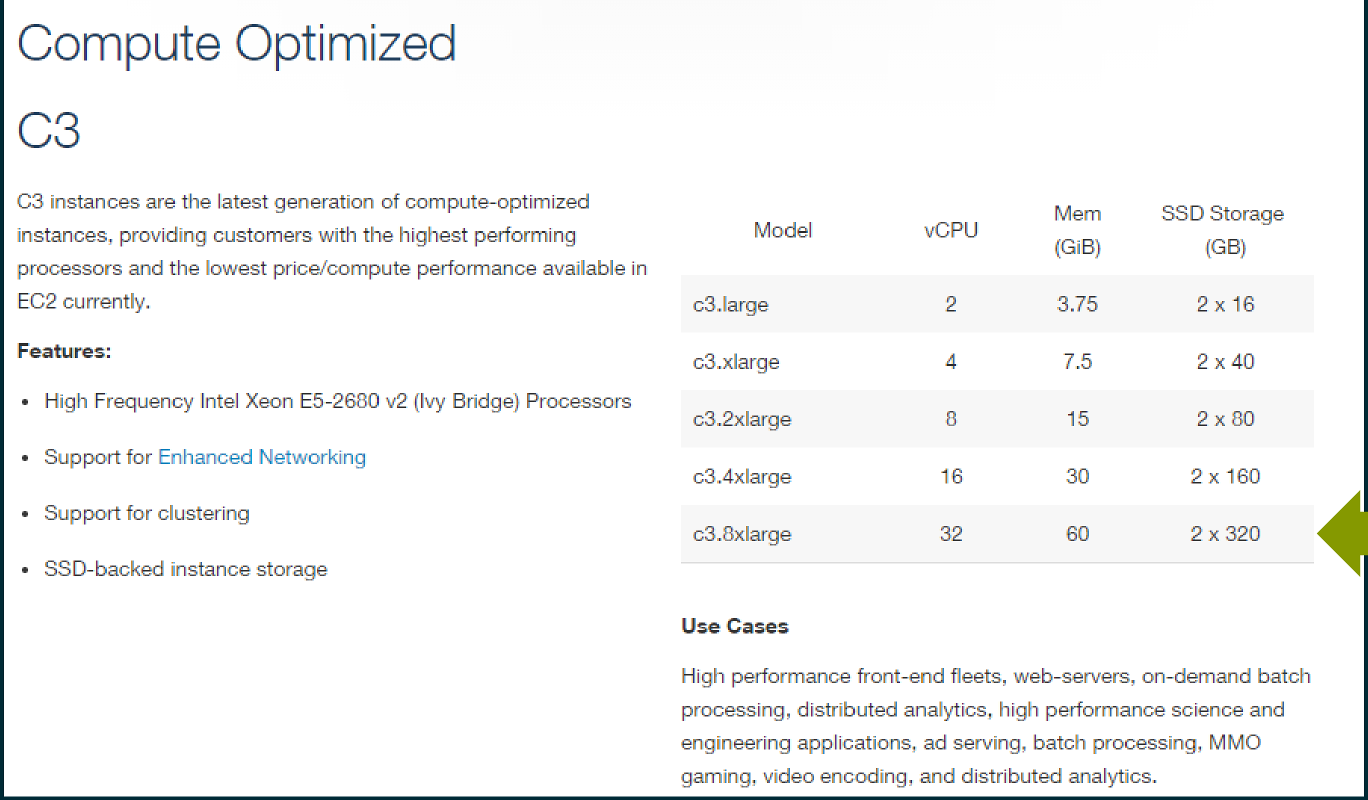
DRAM CPU Disk

Underutilization

Standard ratios



Virtualization



Compute Optimized

C3

C3 instances are the latest generation of compute-optimized instances, providing customers with the highest performing processors and the lowest price/compute performance available in EC2 currently.

Features:

- High Frequency Intel Xeon E5-2680 v2 (Ivy Bridge) Processors
- Support for Enhanced Networking
- Support for clustering
- SSD-backed instance storage

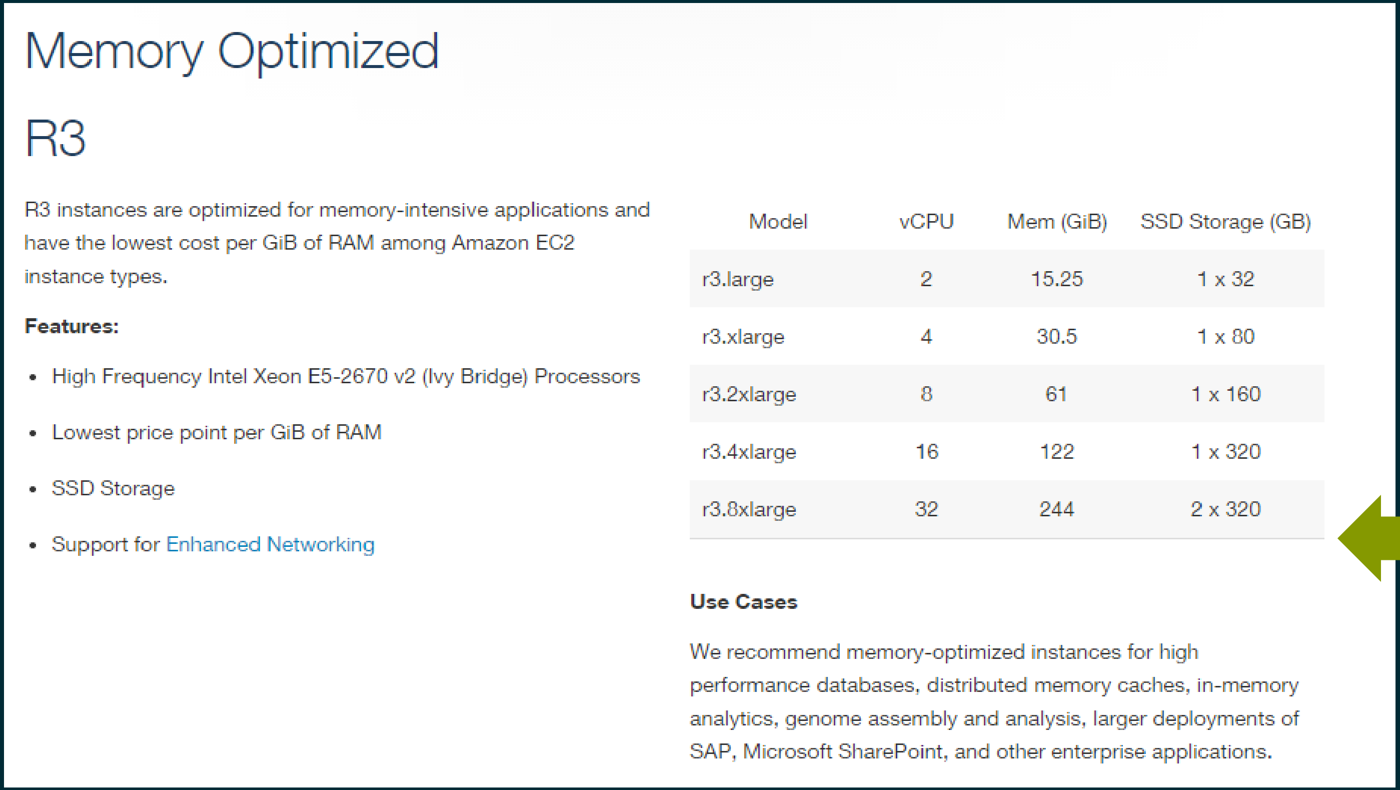
Model	vCPU	Mem (GiB)	SSD Storage (GB)
c3.large	2	3.75	2 x 16
c3.xlarge	4	7.5	2 x 40
c3.2xlarge	8	15	2 x 80
c3.4xlarge	16	30	2 x 160
c3.8xlarge	32	60	2 x 320

Use Cases

High performance front-end fleets, web-servers, on-demand batch processing, distributed analytics, high performance science and engineering applications, ad serving, batch processing, MMO gaming, video encoding, and distributed analytics.

60GB

Virtualization



Memory Optimized

R3

R3 instances are optimized for memory-intensive applications and have the lowest cost per GiB of RAM among Amazon EC2 instance types.

Features:

- High Frequency Intel Xeon E5-2670 v2 (Ivy Bridge) Processors
- Lowest price point per GiB of RAM
- SSD Storage
- Support for Enhanced Networking

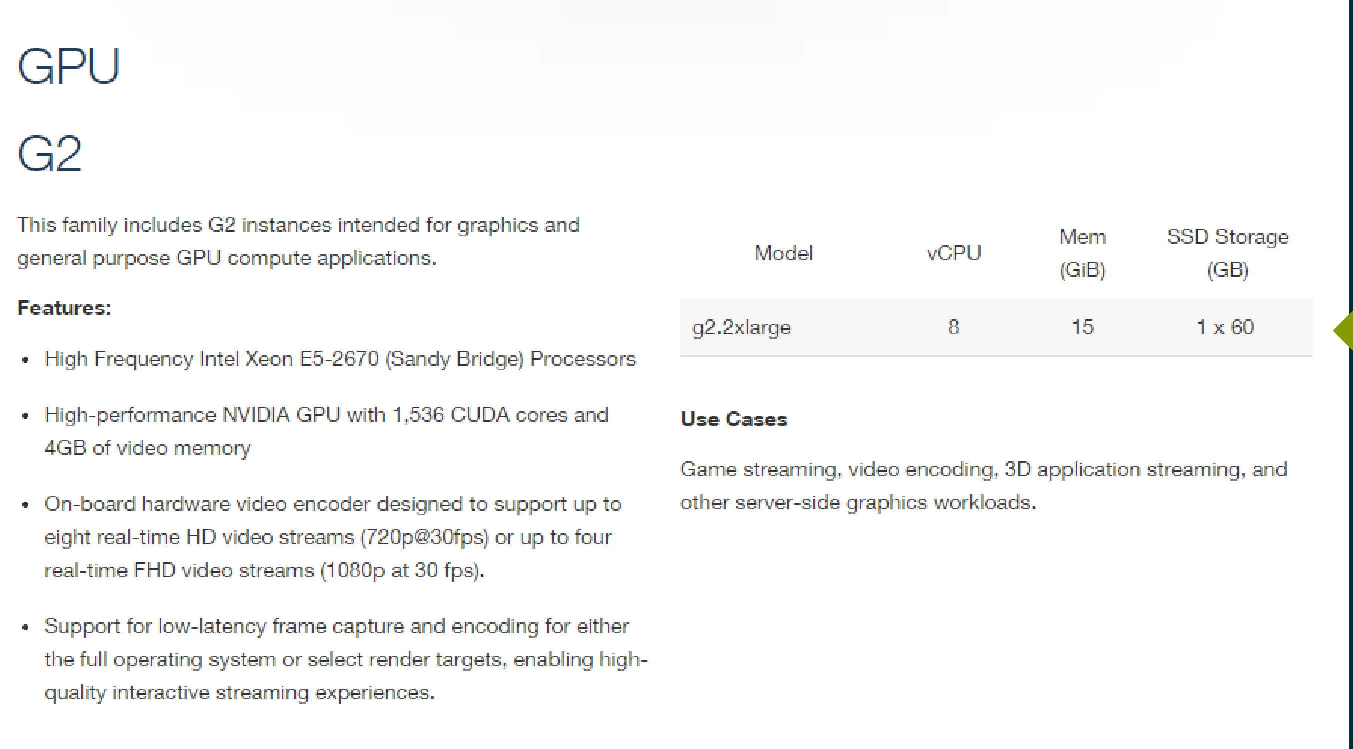
Model	vCPU	Mem (GiB)	SSD Storage (GB)
r3.large	2	15.25	1 x 32
r3.xlarge	4	30.5	1 x 80
r3.2xlarge	8	61	1 x 160
r3.4xlarge	16	122	1 x 320
r3.8xlarge	32	244	2 x 320

Use Cases

We recommend memory-optimized instances for high performance databases, distributed memory caches, in-memory analytics, genome assembly and analysis, larger deployments of SAP, Microsoft SharePoint, and other enterprise applications.

32 CPU

Virtualization



GPU

G2

This family includes G2 instances intended for graphics and general purpose GPU compute applications.

Features:

- High Frequency Intel Xeon E5-2670 (Sandy Bridge) Processors
- High-performance NVIDIA GPU with 1,536 CUDA cores and 4GB of video memory
- On-board hardware video encoder designed to support up to eight real-time HD video streams (720p@30fps) or up to four real-time FHD video streams (1080p at 30 fps).
- Support for low-latency frame capture and encoding for either the full operating system or select render targets, enabling high-quality interactive streaming experiences.

Model	vCPU	Mem (GiB)	SSD Storage (GB)
g2.2xlarge	8	15	1 x 60

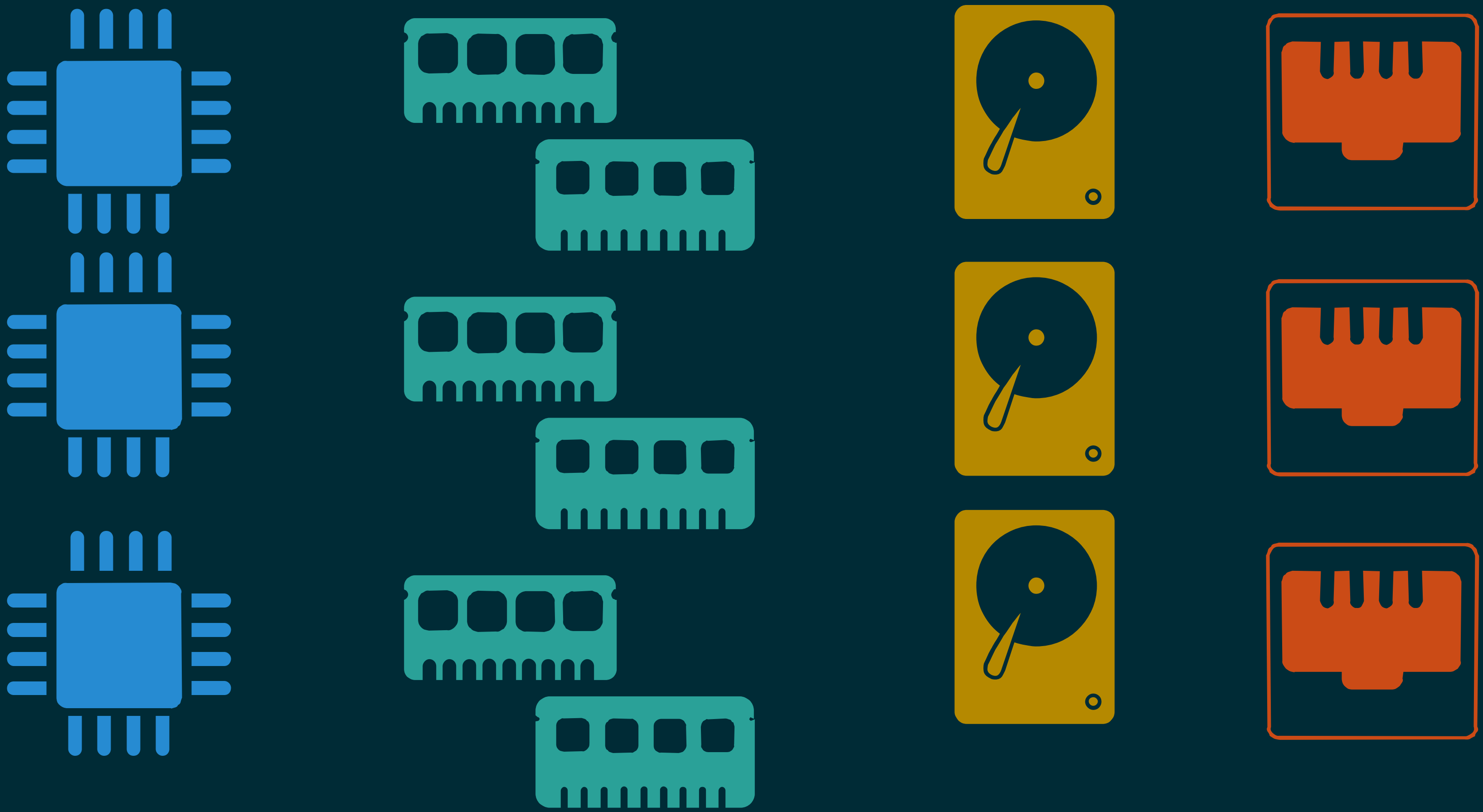
Use Cases

Game streaming, video encoding, 3D application streaming, and other server-side graphics workloads.

8 CPU

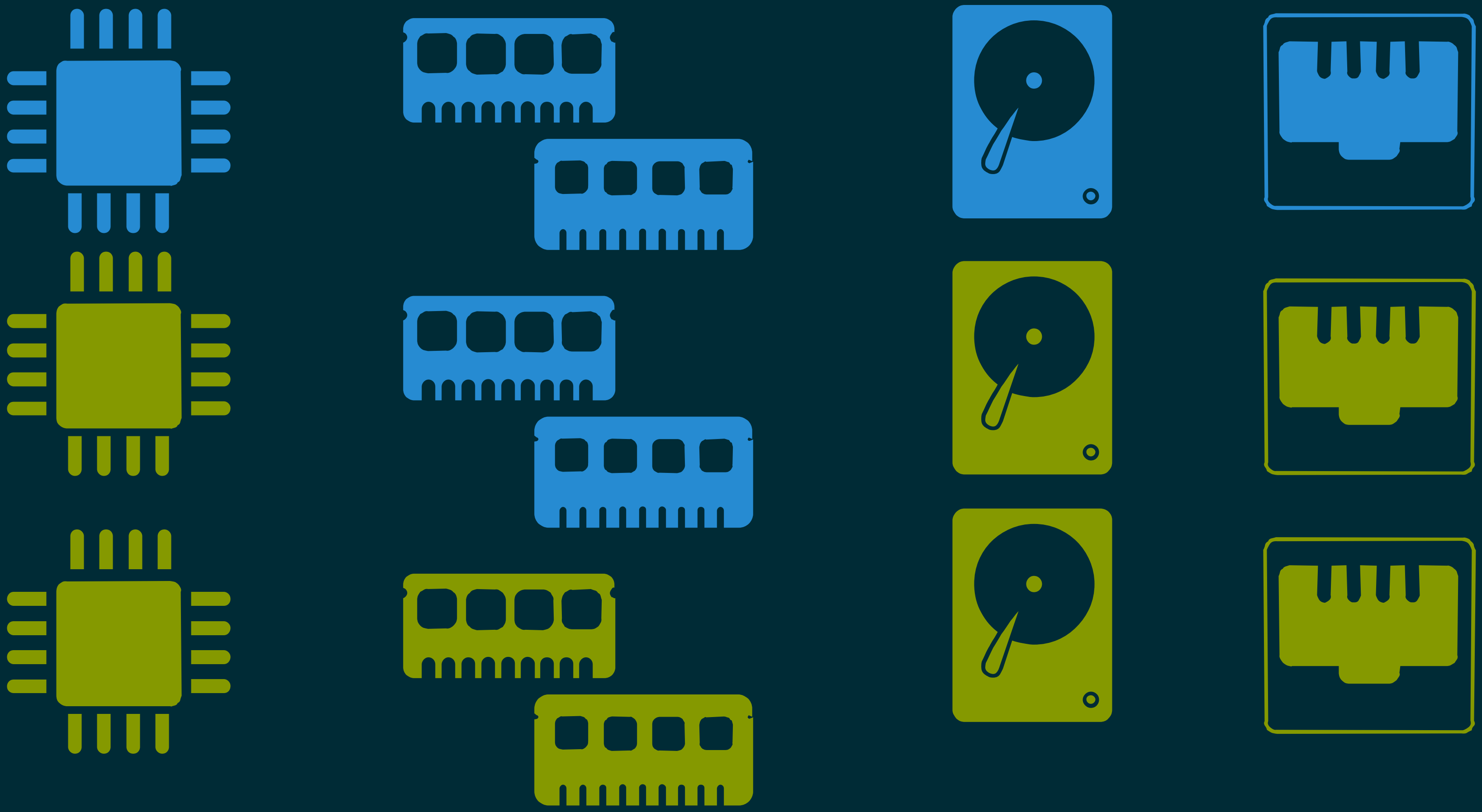
15 GB

Disaggregation 〉 Intro



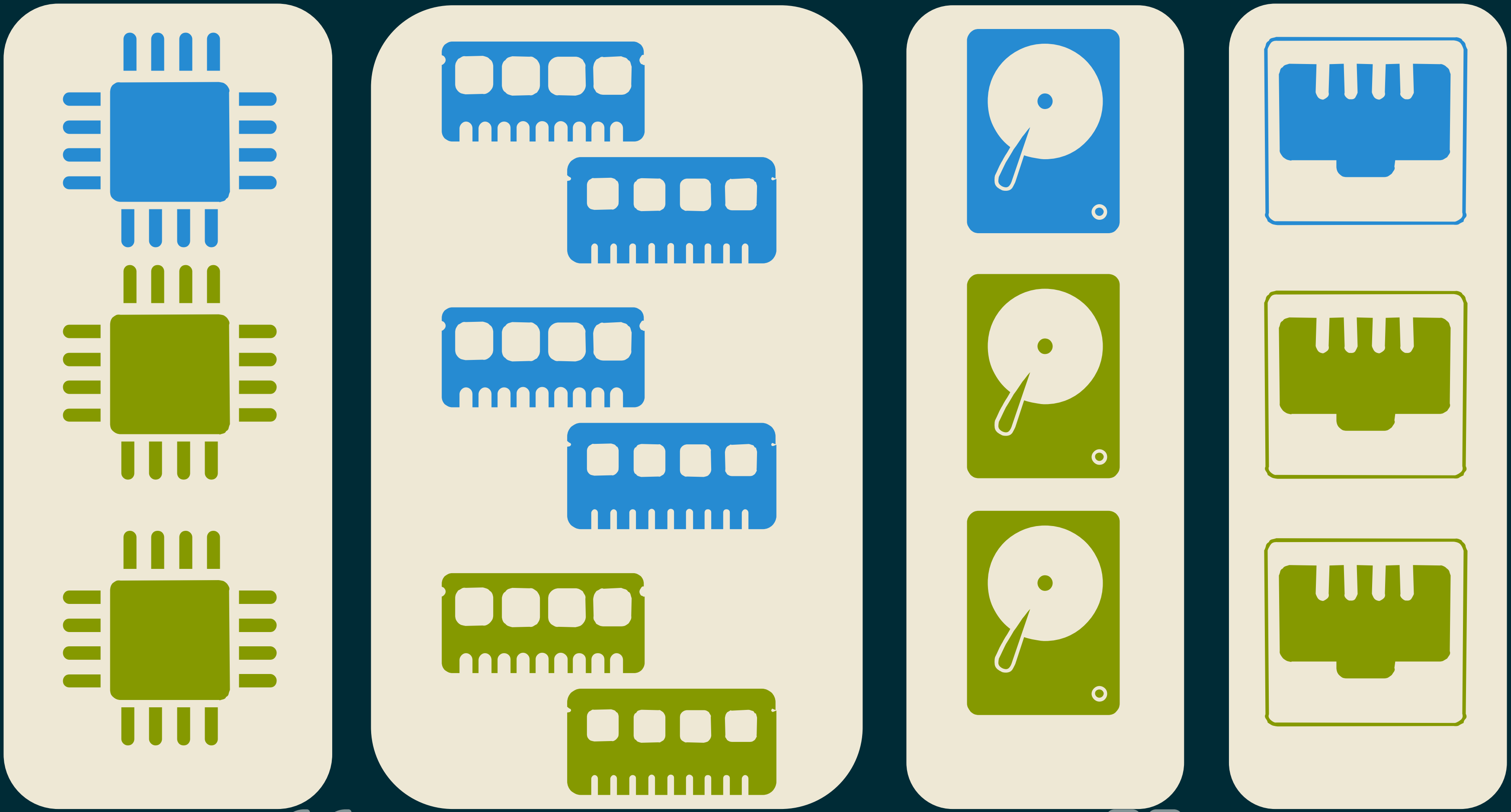
> “Disaggregation”

Disaggregation > Intro



> “Disaggregation”

Disaggregation 〉 Intro



“Disaggregation”

Disaggregation 〉 History

Network Support for Resource Disaggregation in Next-Generation Datacenters [Han et al. '13], HOTNETS

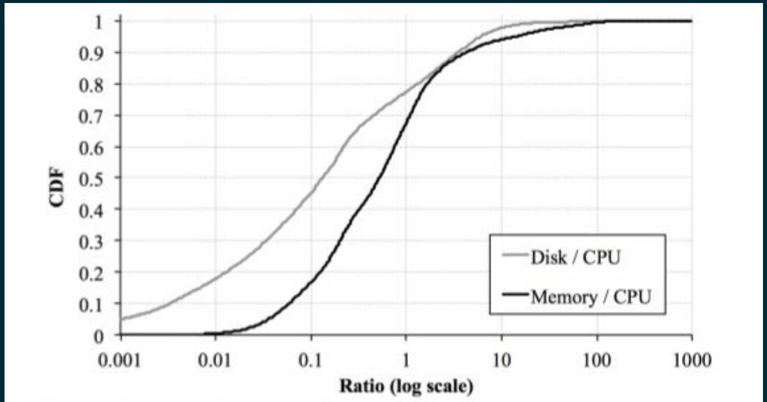
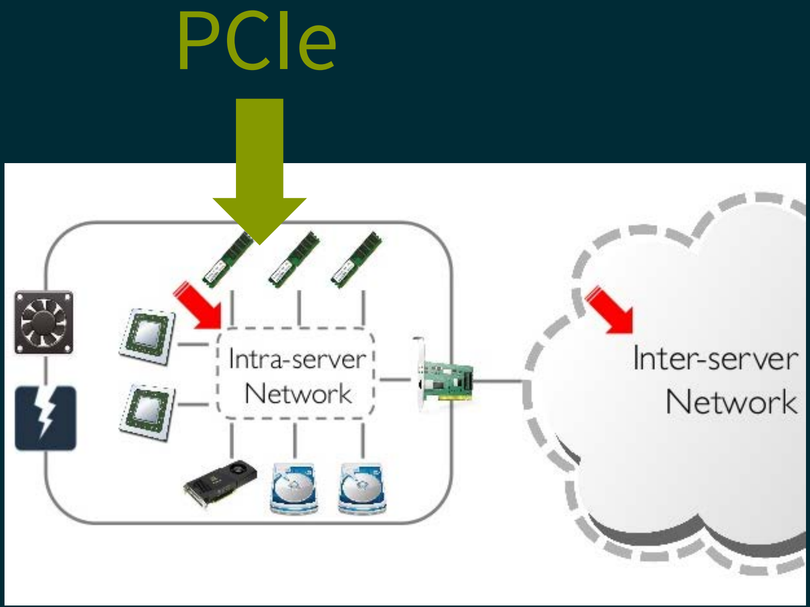


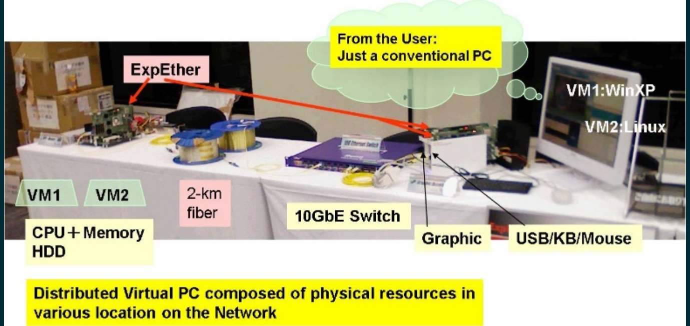
Figure 1: Distribution of disk/memory capacity demand to CPU usage ratio for tasks in Google's datacenter.



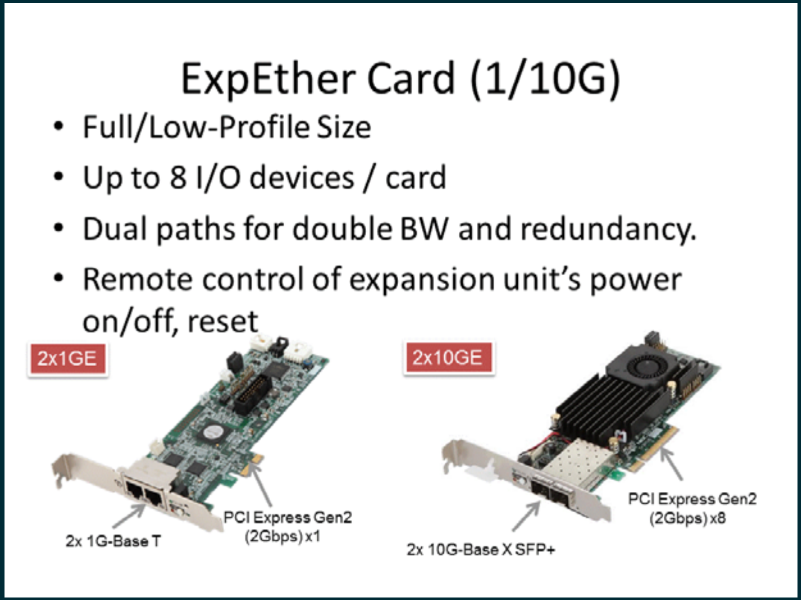
Network	Rack Scale
Resource	All
Topology	Multi Master

Disaggregation 〉 History

End-to-End Adaptive Packet Aggregation for High-Throughput I/O Bus [Suzuki et al. ’13], HOTInterconnects



Distributed Virtual PC composed of physical resources in various location on the Network



ExpEther Card (1/10G)

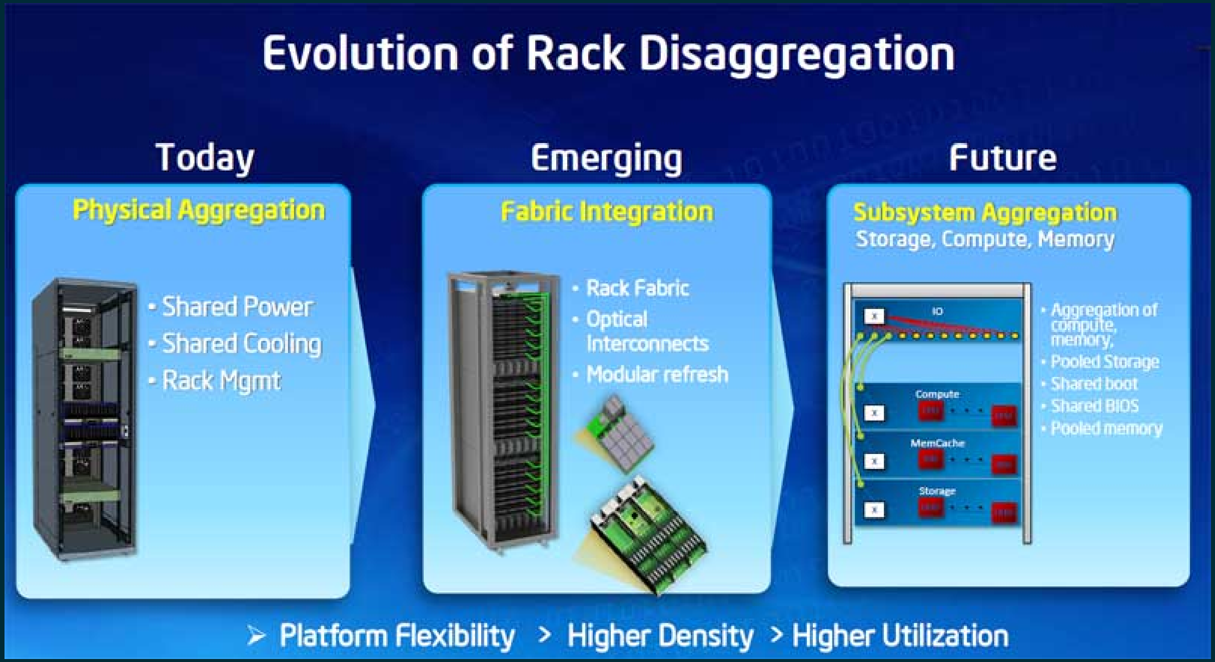
- Full/Low-Profile Size
- Up to 8 I/O devices / card
- Dual paths for double BW and redundancy.
- Remote control of expansion unit’s power on/off, reset

2x1GE
2x 1G-Base T
PCI Express Gen2 (2Gbps) x1
2x10GE
2x 10G-Base X SFP+
PCI Express Gen2 (2Gbps) x8

Network	Local Area
Resource	I/O
Topology	Single Master

Disaggregation > History

Intel Rack Scale Infrastructure



Evolution of Rack Disaggregation

Today

Physical Aggregation

- Shared Power
- Shared Cooling
- Rack Mgmt

Emerging

Fabric Integration

- Rack Fabric
- Optical Interconnects
- Modular refresh

Future

Subsystem Aggregation
Storage, Compute, Memory

- Aggregation of compute, memory,
- Pooled Storage
- Shared boot
- Shared BIOS
- Pooled memory

IO
Compute
MemCache
Storage

➢ Platform Flexibility > Higher Density > Higher Utilization

Network	Rack Scale
Resource	All
Topology	Multi Master

Disaggregation > Summary

- Most designs use a rack-scale network, disaggregate I/O only, and are multi-root.

- Are these designs really disaggregated, is the size of the aggregate now a rack?

Problem

- Aggregation results in inefficient fixed-ratio physical provisioning
- Disaggregation can solve the problem, but drastically increases network requirements

Our Approach

- Disaggregate using local area network, giving scalability
- Circuit switching can help manage latency and bandwidth requirements
- Use a peer-to-peer design, where each peer can access other peer’s resources

Disaggregation 〉 Proposal

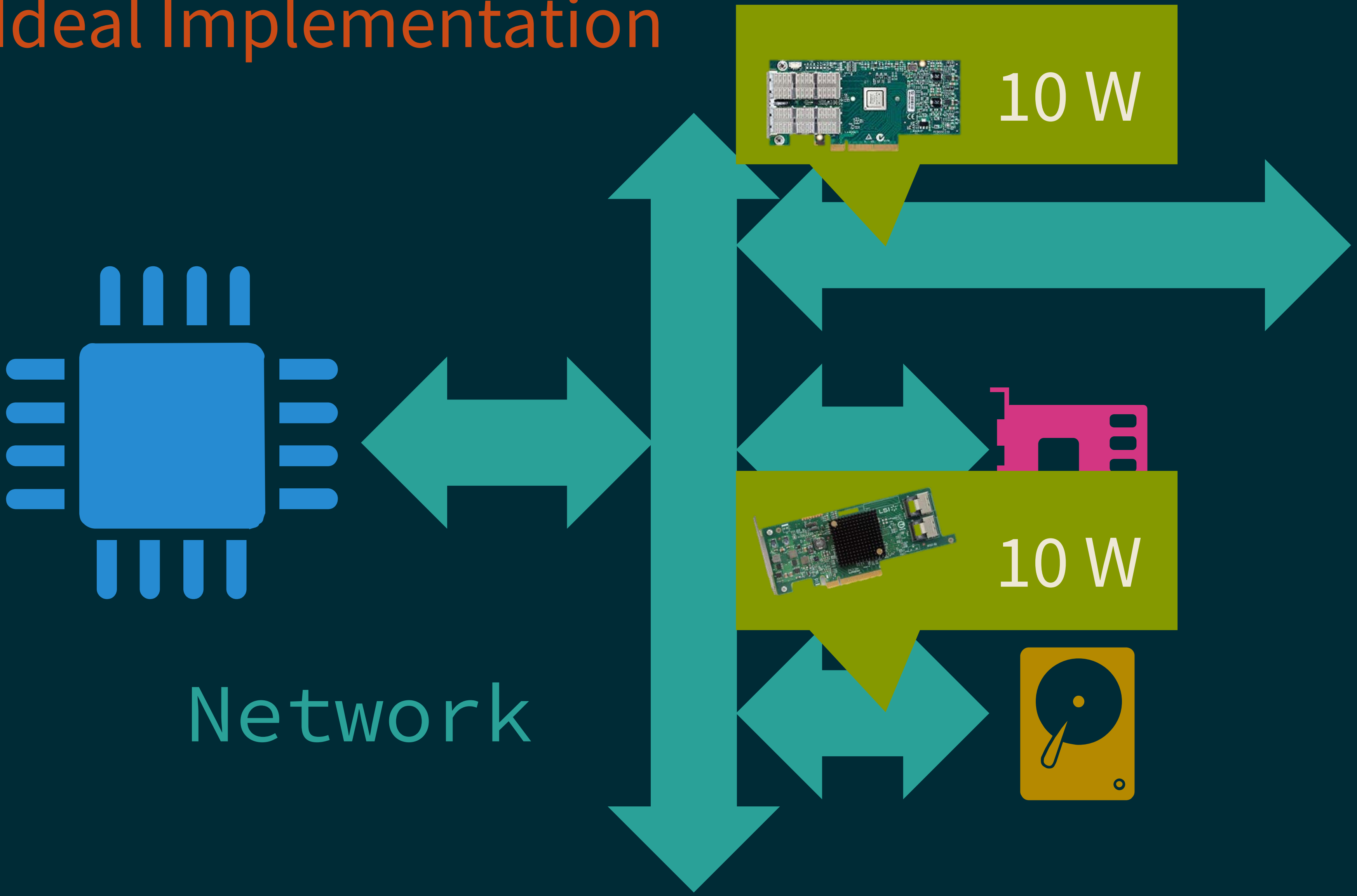
Applications

- Goal is to build a system where the application drives resources, rather than the other way around.

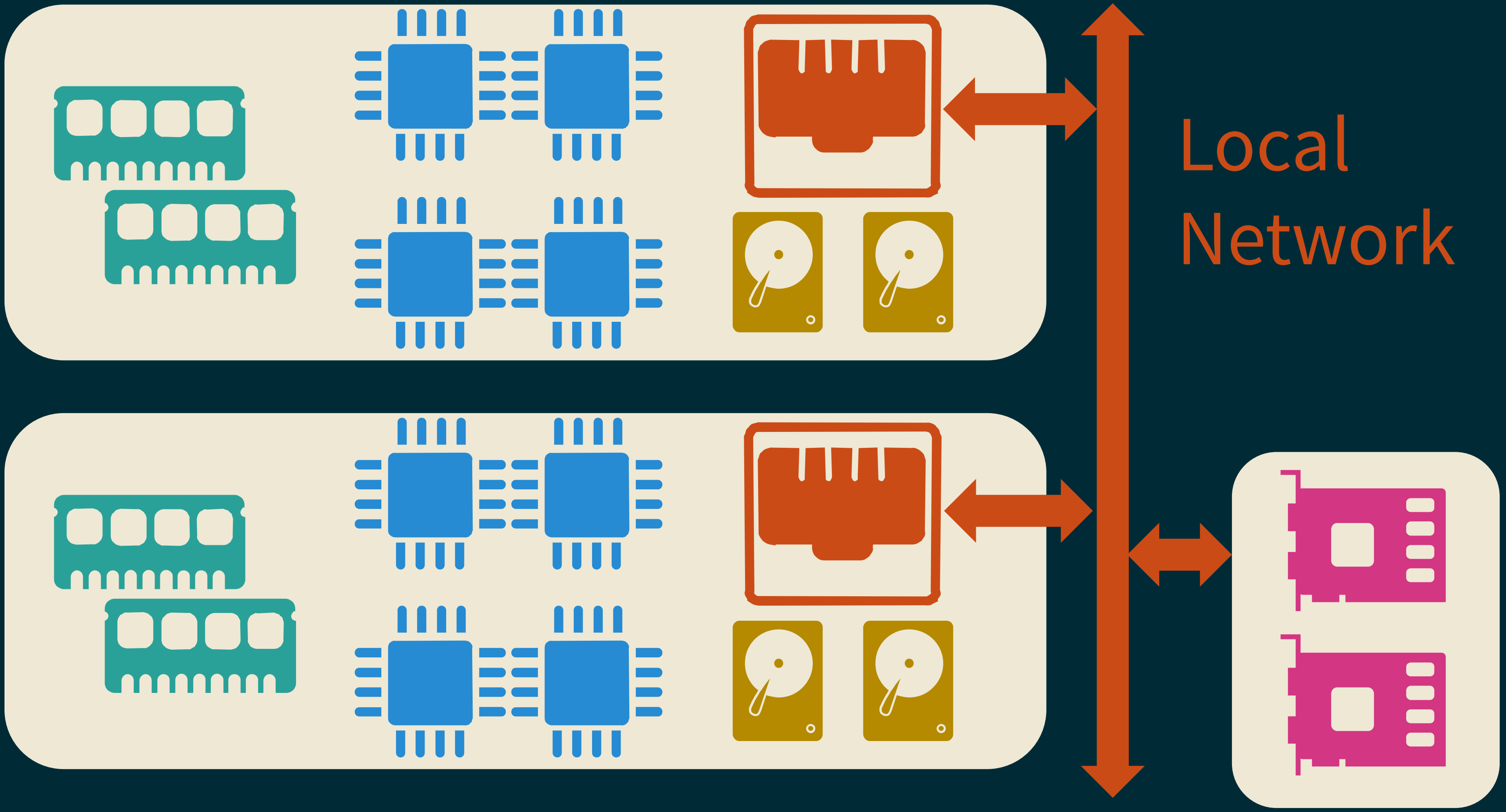
- Changes how computers are fundamentally built, making them more energy-efficient and cost-effective

Our System > Implementation

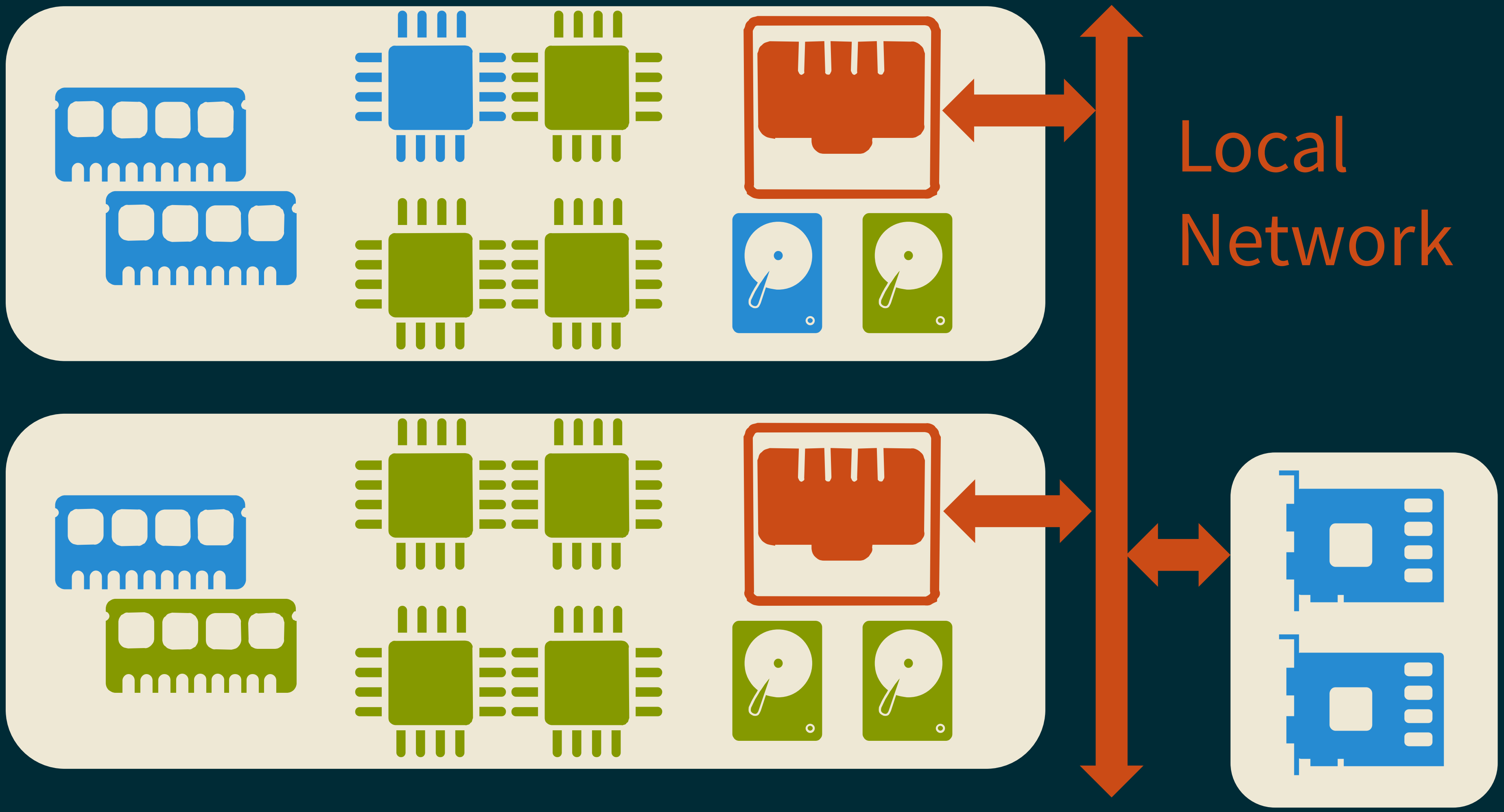
Ideal Implementation



Our System 〉 Architecture



Our System 〉 Architecture



Conclusion

- Aggregation leads to underutilization and inefficiency
- Disaggregation can increase utilization, but taxes the network
- Existing designs use rack-scale networks, limiting scalability
- Our design uses commodity network and servers
- Everything is network has wider applications