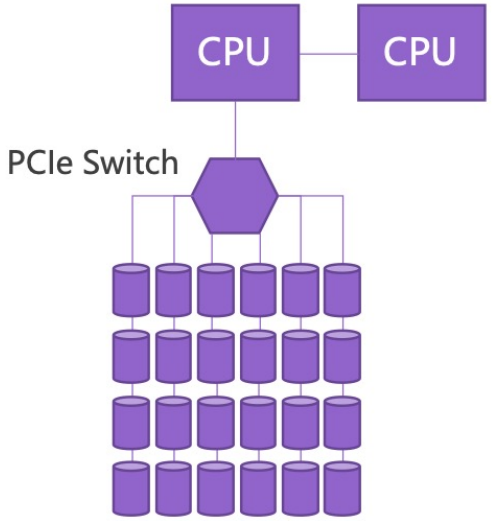
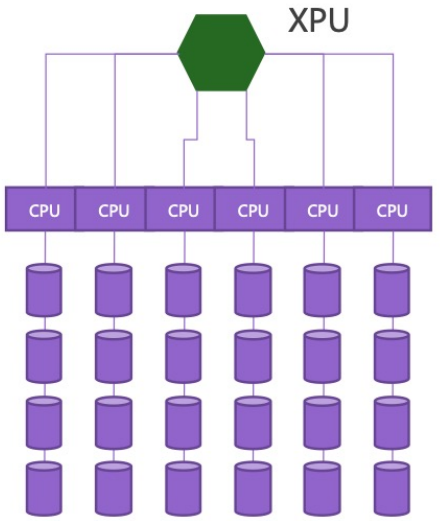
PCIe Net enable Object Storage at Edge

Traditional Storage
Elephant



- High cost
- High power
- Not cloud native

Micro-server Architecture
Ants



- Low cost
- Low power
- Cloud native

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

294bn

billion emails are sent

Radicati Group

320bn emails to be sent each day by 2021

306bn emails to be sent each day by 2020

3.9bn

people use emails

4TB

of data produced by a connected car

Intel

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1,000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1,000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1,000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1,000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1,000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1,000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1,000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

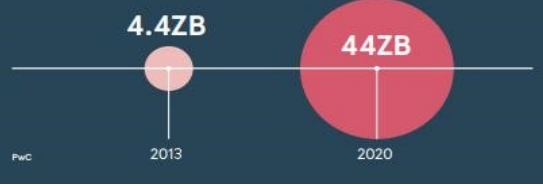
*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

65bn

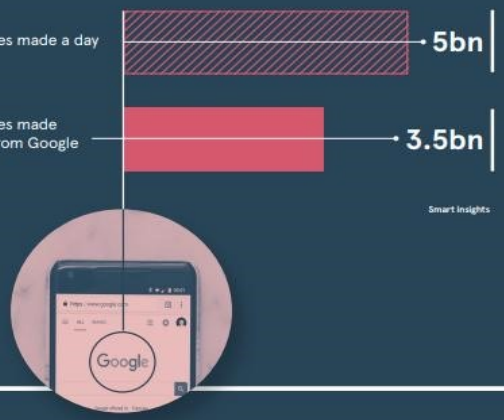
messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

ACCUMULATED DIGITAL UNIVERSE OF DATA



PwC



Smart Insights

463EB

of data will be created every day by 2025

IDC

95m

photos and videos are shared on Instagram

Instagram Business

28PB

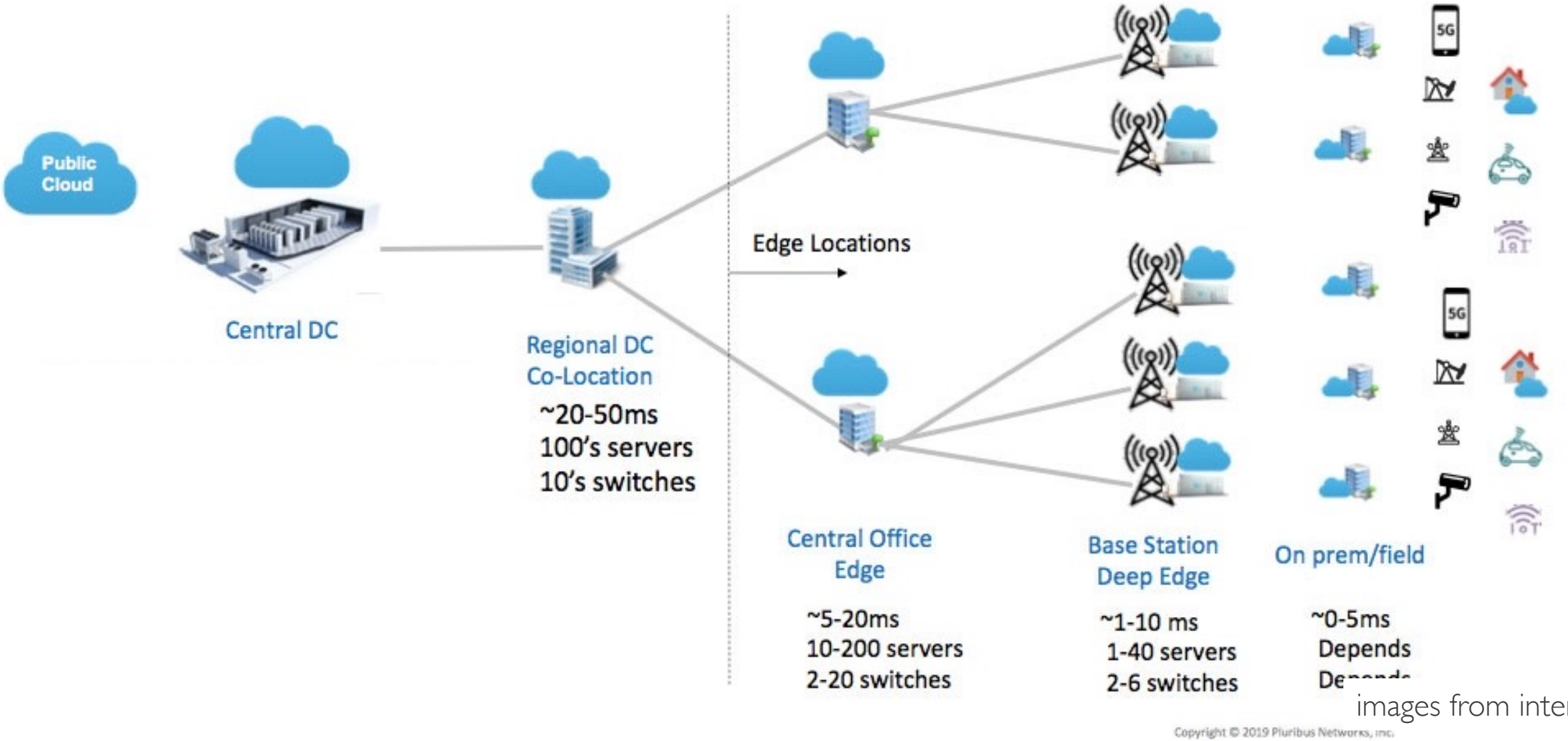
to be generated from wearable devices by 2020

Statista

https://cn.weforum.org/agenda/2019/06/mei-tian-neng-chan-sheng-duo-shao-shu-ju/

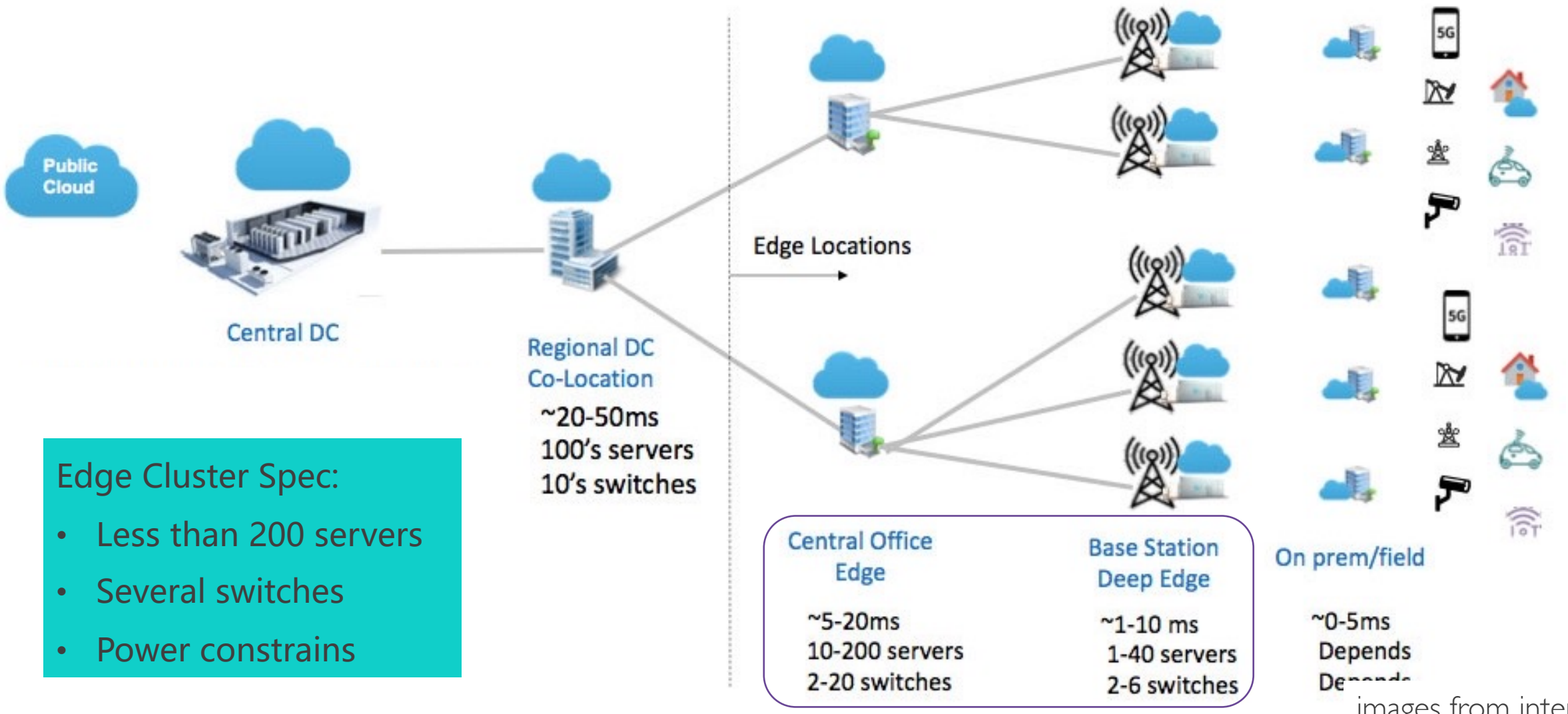
RACONTEUR

Data Flow from Cloud to Deep Edge



images from internet

Edge Cluster Scale-in

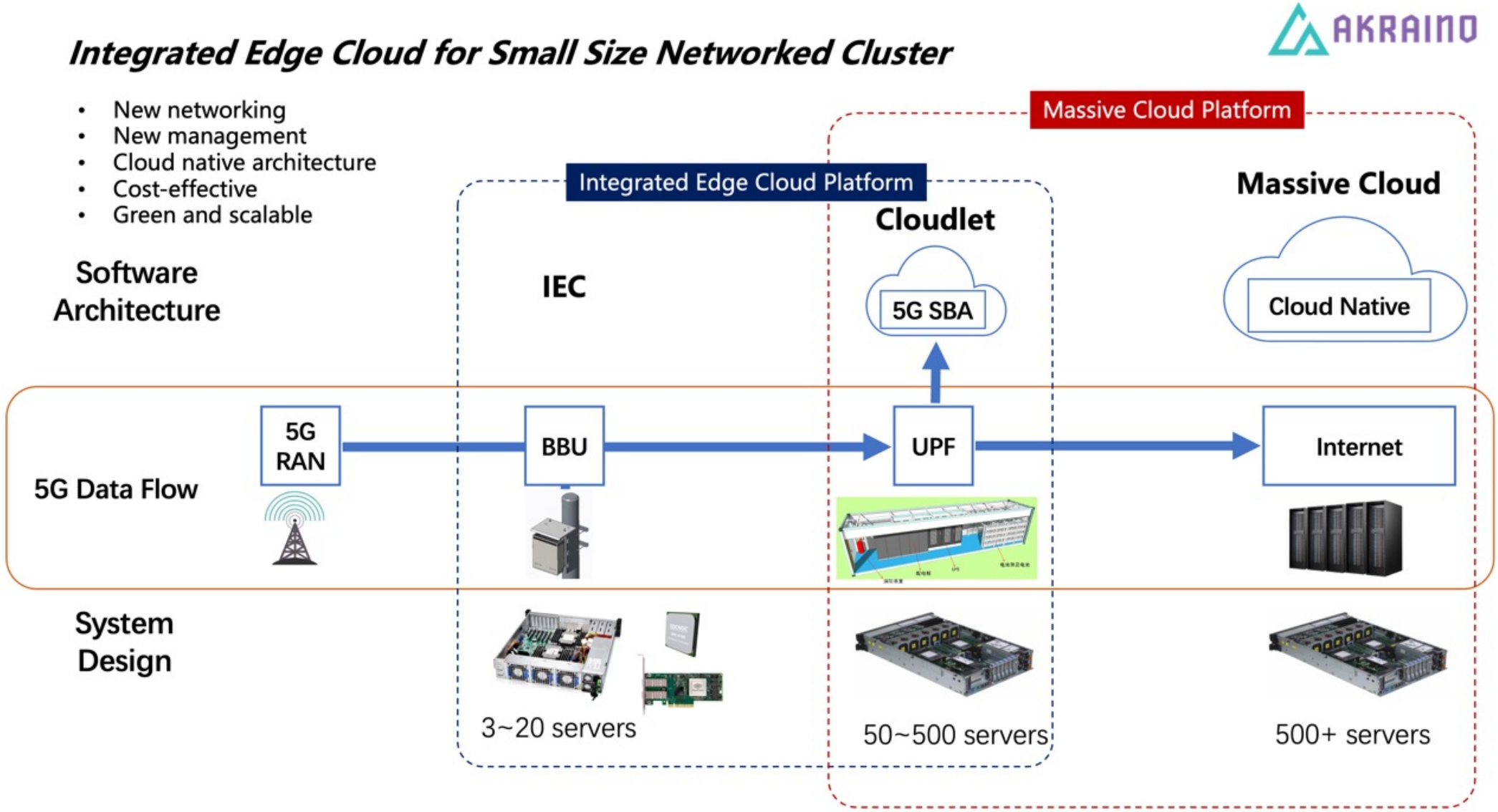


Edge Cluster Spec:

- Less than 200 servers
- Several switches
- Power constrains

images from internet

Integrated Edge Cloud (IEC) - Akraino Type 5



Scale-in Cluster at Edge

In contrast to centralized data centers, the networking in the compact integrated edge cloud needs to be reconsidered, as the networking challenge shifts from solving for the *scale-out* to a massive number of connected servers to enabling the *scale-in* for connecting a small number of servers in an edge location. The time-honored methods of increasing the *port density* in a switch or utilizing the high throughput NIC won't work in a small scale cluster with less than 32 servers. Hence a novel way to rebuild the networking architecture for the integrated edge cloud is needed not only in terms of cost but also, and even more importantly, due to energy constraints, given the expected large number of deployments of the integrated edge cloud sites.

Compact Networking for Small Cluster

"Due to the dynamic nature of edge deployments, it is critical that networks are able to adapt to current operational context in order to optimize performance and uptime. "

and Industrial worlds, many systems communicate over legacy local area networks such as 4-20mA current loops, serial and CAN Bus, as well as modern low-power wireless technologies such as Bluetooth and LoRa. IoT gateways serve the function of converting these transports into IP traffic.

In terms of application-level protocols, there is a wide array of choices to contend with when developing edge solutions. While there are tens that matter in the IT world (e.g. REST, MQTT), there are literally hundreds if not thousands in the OT/ Industrial world, with examples including Modbus, BACnet, PROFINET and Ether-CAT. Edge solutions often have to comprehend a blend of these application-level protocols and LF Edge projects like EdgeX Foundry and Fledge are focused on simplifying data flow in heterogeneous environments.

Due to the dynamic nature of edge deployments, it is critical that networks are able to adapt to current operational context in order to optimize performance and uptime. This can include dynamically switching between available connections.

Project Contributions for Edge Connectivity

The LF Edge projects are addressing connectivity needs both at the application level for protocol normalization to facilitate IoT interoperability and transport level in terms of network virtualization and optimization. The following are examples of each project's focus in the area of connectivity.

AKRAINO

Akraino blueprints can provide an end to end EdgeStack to support Virtualized Network Elements (NFVI) per Open-RAN (O-RAN) requirements. The Akranio project advocated collaborating with O-RAN Alliance's specification workgroup 6 (six) respon-sible for cloud specifications, to align with and publish multiple blueprints to support various RAN use cases for Radio Edge cloud, including ORAN-Software Community's Near-RT RIC software, Network Cloud with RS-IOV or OVS-DPDK, Integrated Cloud Native, Kubernetes Native Infrastructure provider Access edge and more.

There are a number of blueprints that provide interesting examples for edge connectivity that can be applied in different parts of the edge ecosystem. As an example, the Network Cloud with Tungsten Fabric (TF) blueprint provides a fully distributed networking stack based on a microservices architecture, implementing a distributed networking framework for Edge computing. TF SDN Controller provides seamless and full integration between different types of workloads VNFs, CNFs and PNFs using a common networking stack integrated with different orchestration platforms like OpenStack and Kubernetes. The TF SDN Controller works as single entity running at the core, distributed core or edge sites, or public cloud (AWS, Azure, GCP or Equinix Metal) and fully integrated with OpenStack Neutron Plugin, Kubernetes CNI, for all types of Edge computing workloads. The solution provides the Tungsten Fabric Kernel vRouter, DPDK vRouter, and support for SR-IOV and SmartNIC.

"The Network Cloud with Tungsten Fabric (TF) blueprint provides a fully distributed networking stack based on a microservices architecture, implementing a distributed networking framework for Edge computing."

Another example of an innovative Akraino blueprint is the Integrated Edge Cloud (IEC) for compact edge with PCIe networking and Cloud-on-Board (CoB). In contrast to centralized data centers, the networking in the compact integrated edge cloud needs to be reconsidered, as the networking challenge shifts from solving for the scale-out to a massive number of connected servers to enabling the scale-in for connecting a small number of servers in an edge location. The time-honored methods of increasing the port density in a switch or utilizing the high throughput NIC won't work in a small scale cluster with less than 32 servers. Hence a novel way to rebuild the networking architecture for the integrated

28

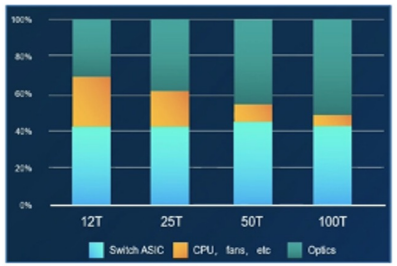
XPU Based Cloud Native Server:

Architecture, Implementation & Applications

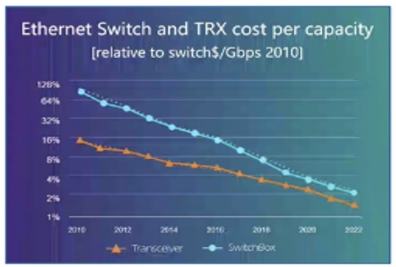
Dr. Fu Li (LEO)

Real Problem Computing Cluster Faced!

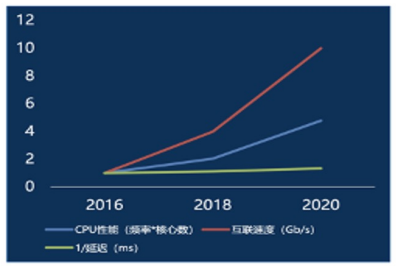
› Optics are too expensive both in ***power*** and ***cost***



光模块功耗占比已经超过50%



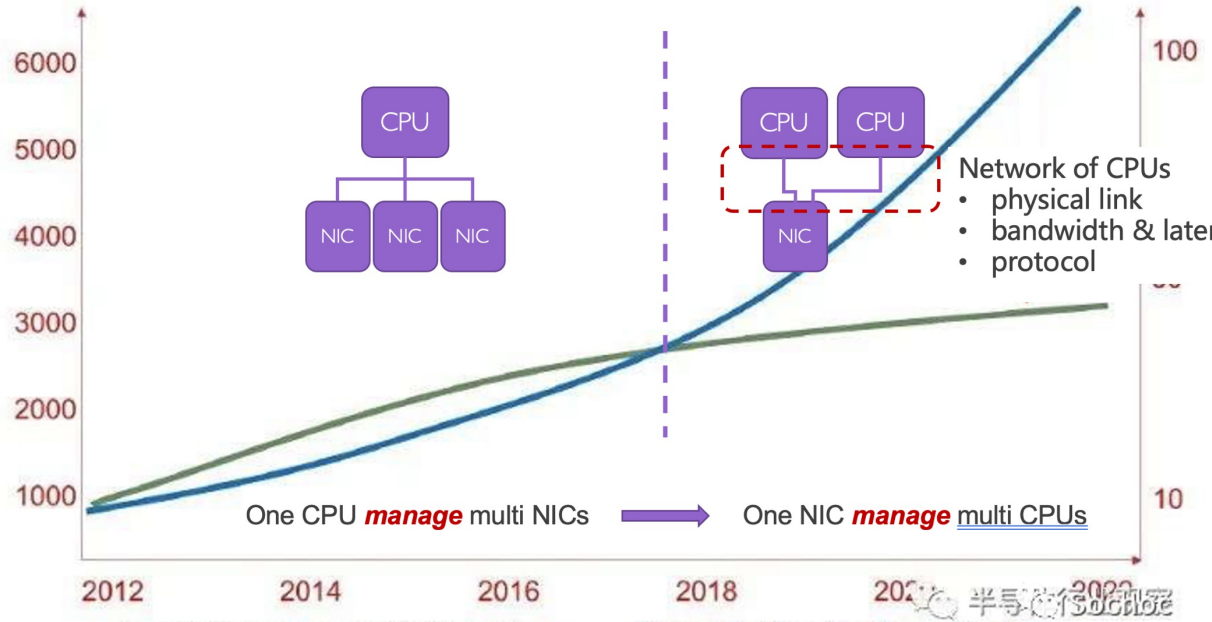
光模块成本已经超过通道成本



延迟降低进展缓慢

Socnoc
SOCNOC

Problem Revisit: Clustered CPU on Board



Data Fabric Landscape

TCP/IP

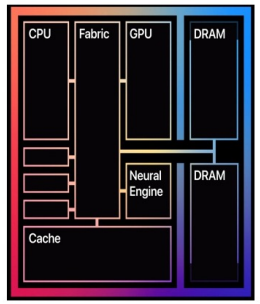
PCIe/CXL
cxl.io/cxl.cache

UCIe

Chiplet

I/O module

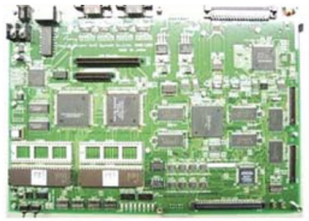
System-in-Package



Interposer Fabric

In-Package Links

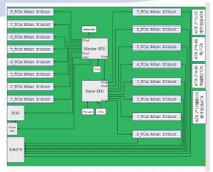
System-on-Board



PCIe Bus based link

NEW

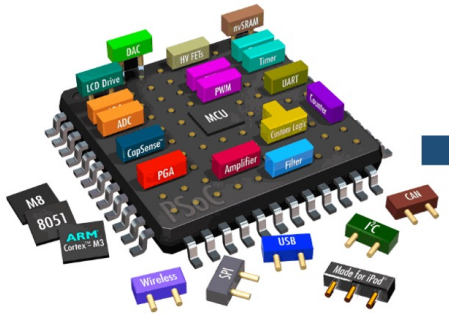
Cloud-on-Board AKRAINO



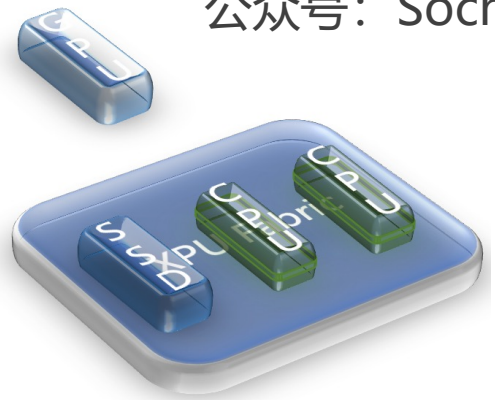
PCIe Net based Fabric

Off-Package Links

XPU: Center of Data Fabric



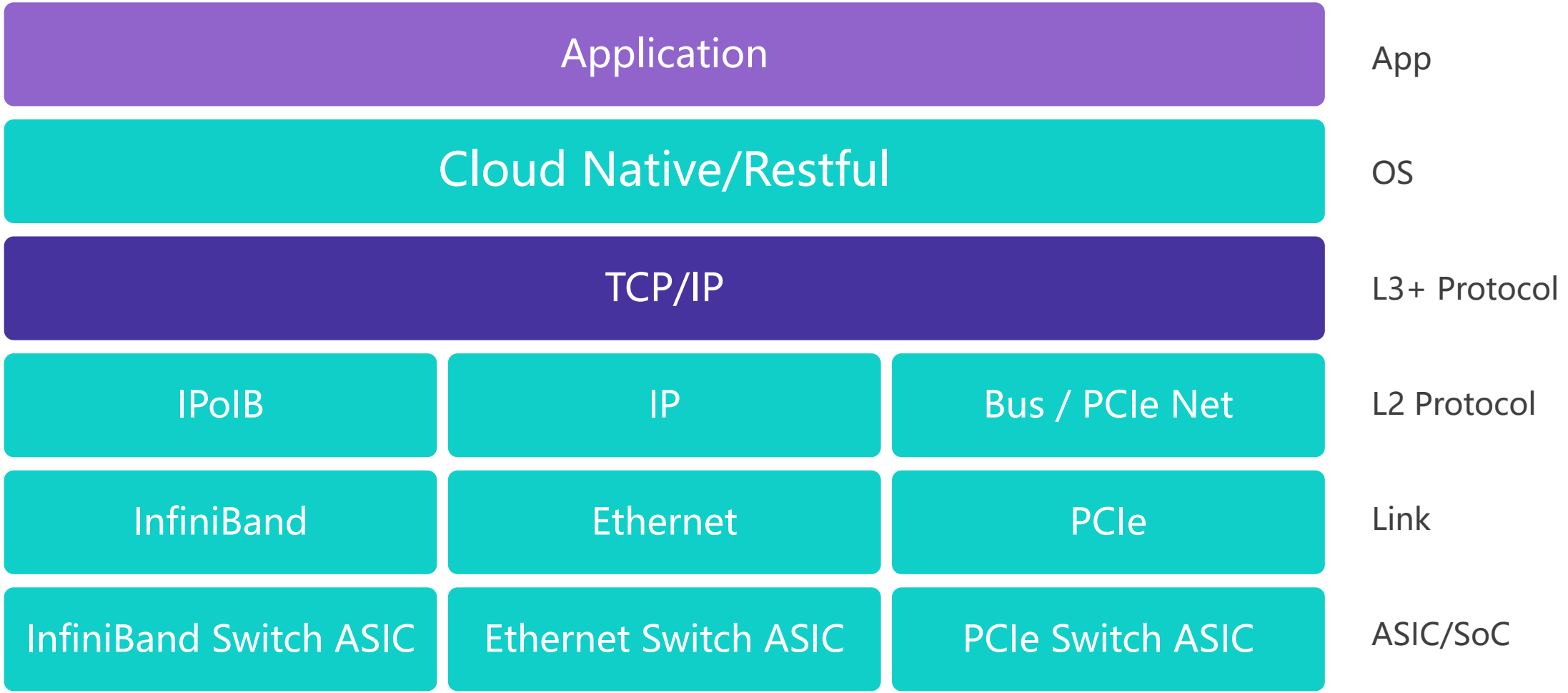
In Package Technology



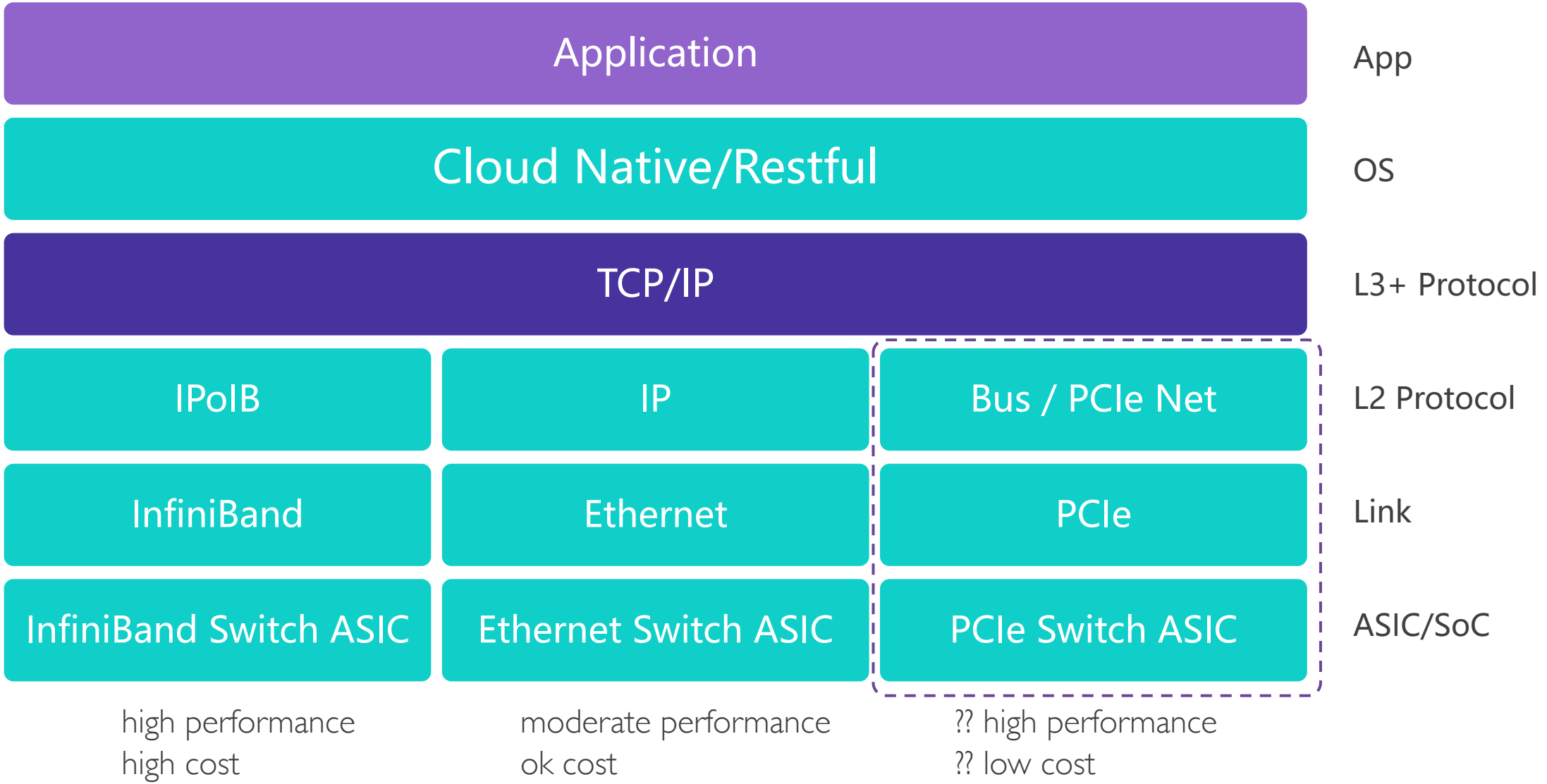
Off Package Technology

公众号：Socnoc

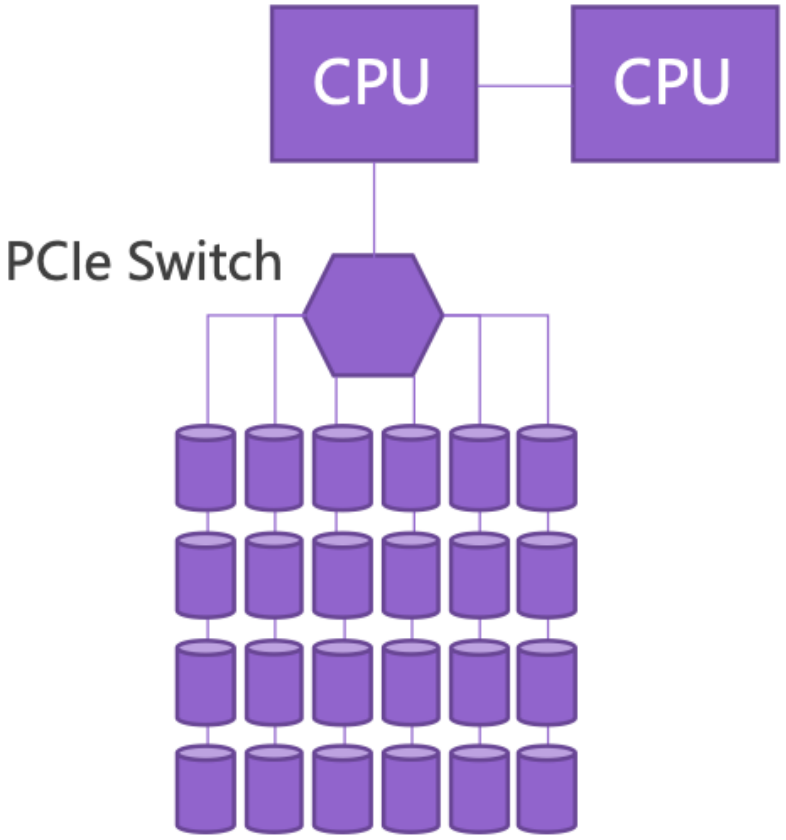
Networking Technologies for Clusters

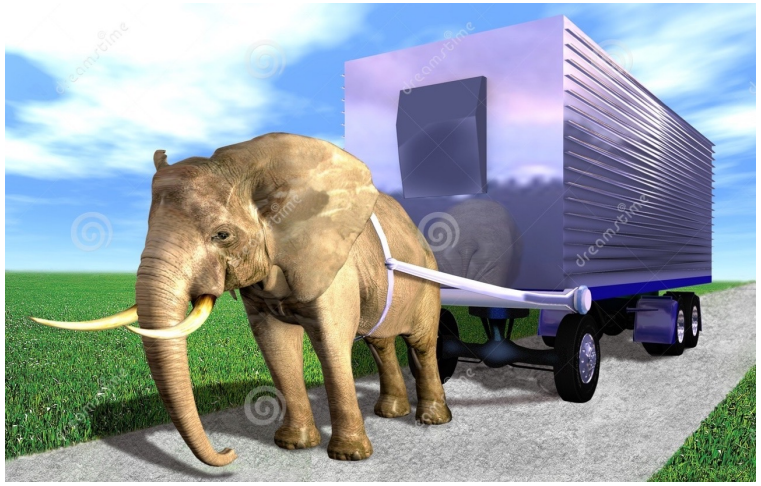


PCIe Net and System Bus



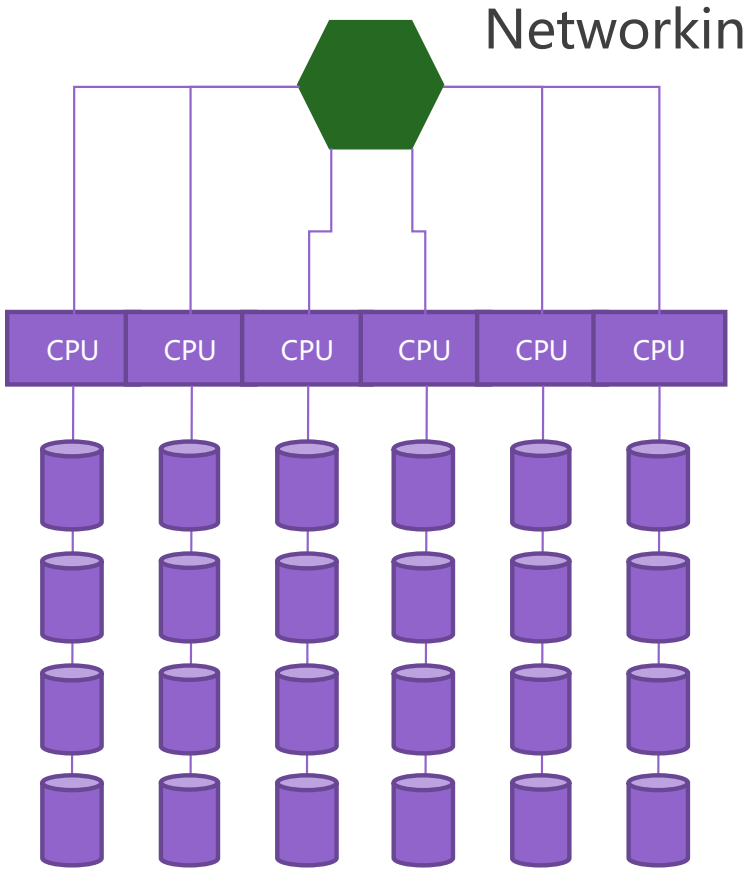
PCIe Fanout Storage System (Elephant)

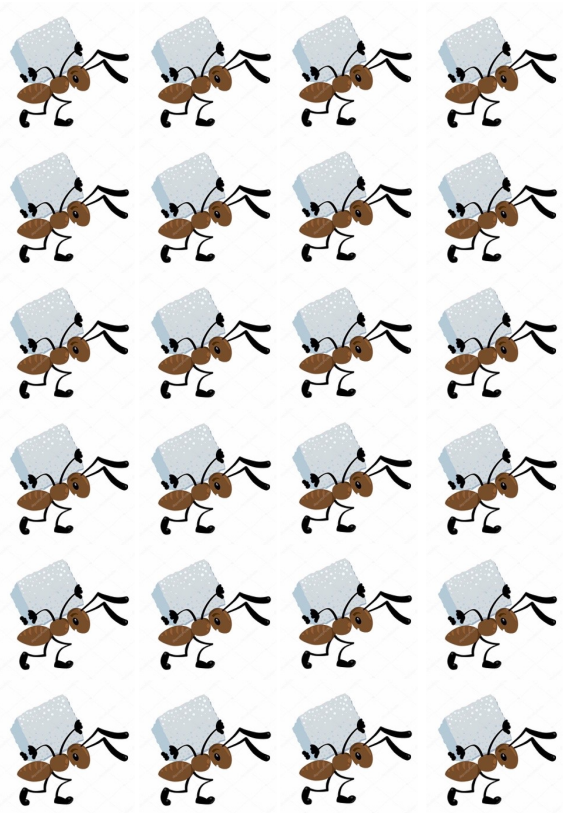




images from internet

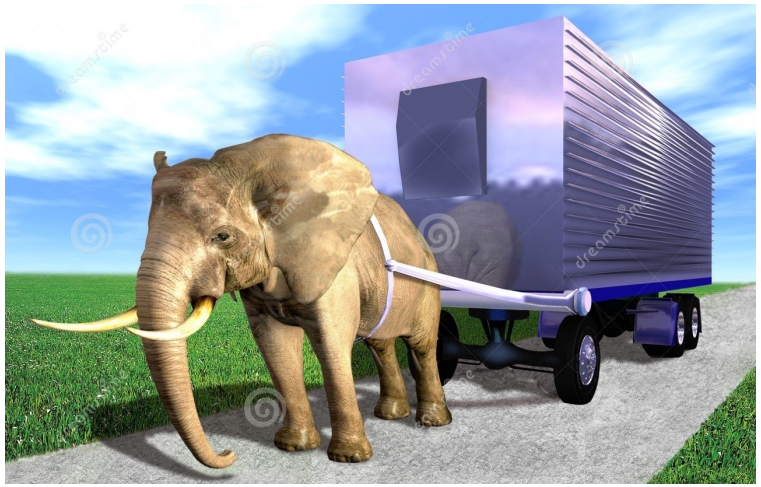
Micro server Storage System (Networked Ants)



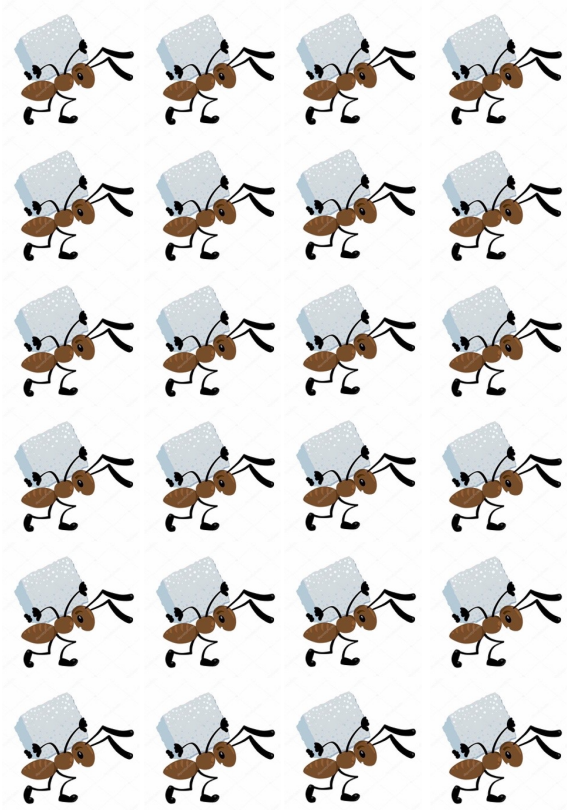


from partner and internet

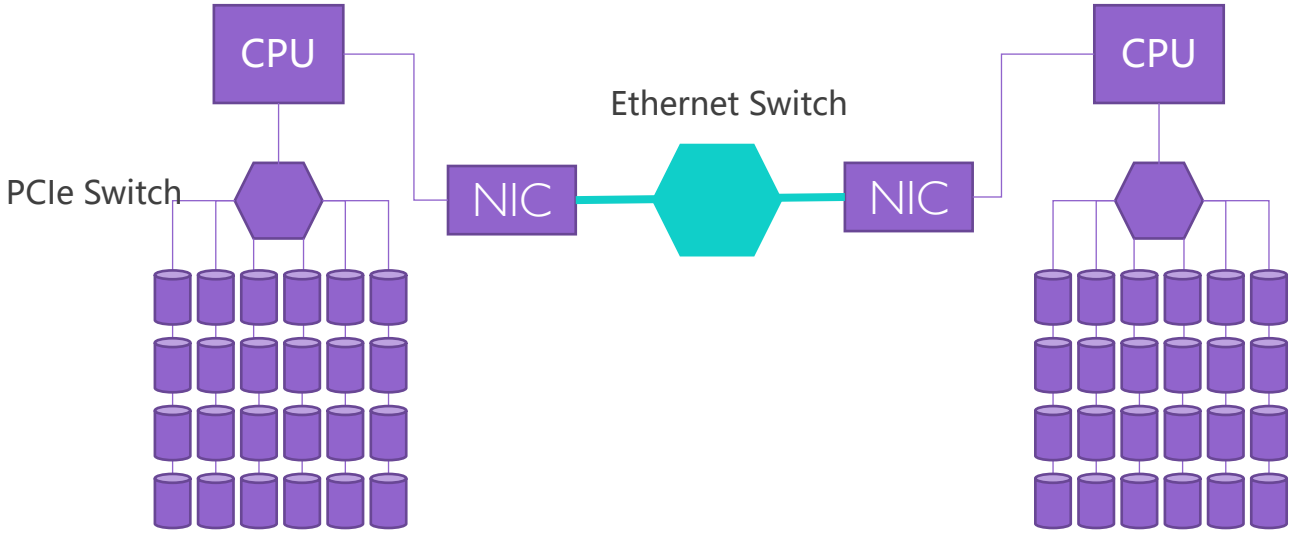
Centralized vs Distributed System



vs

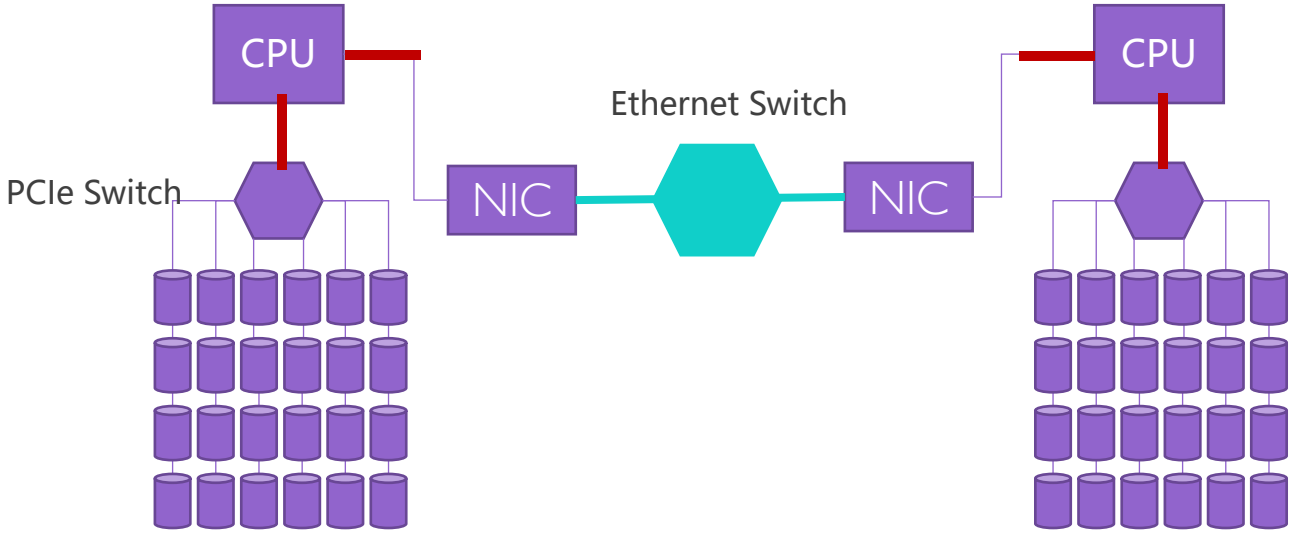


Deep Thinking on System Architecture



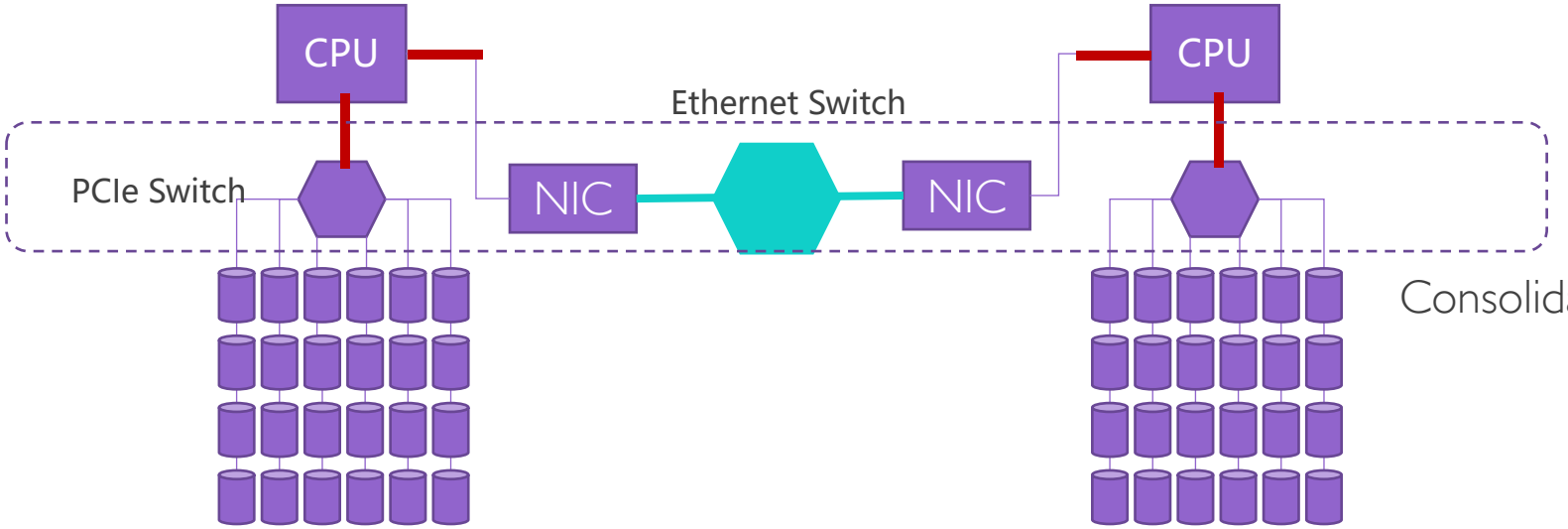
PCIe Switch for Devices, Ethernet Switch for CPUs/Servers

Architecture Bottleneck



	CPU/SoC #	PCIe Ports	Ethernet Switch Ports	Bottleneck
Elephant (HPC server)	2~4	24~36	2~4	Ethernet / CPU
Ants (microserver)	8~24	2-4	8~24	Ethernet Latency

Architecture Innovation?



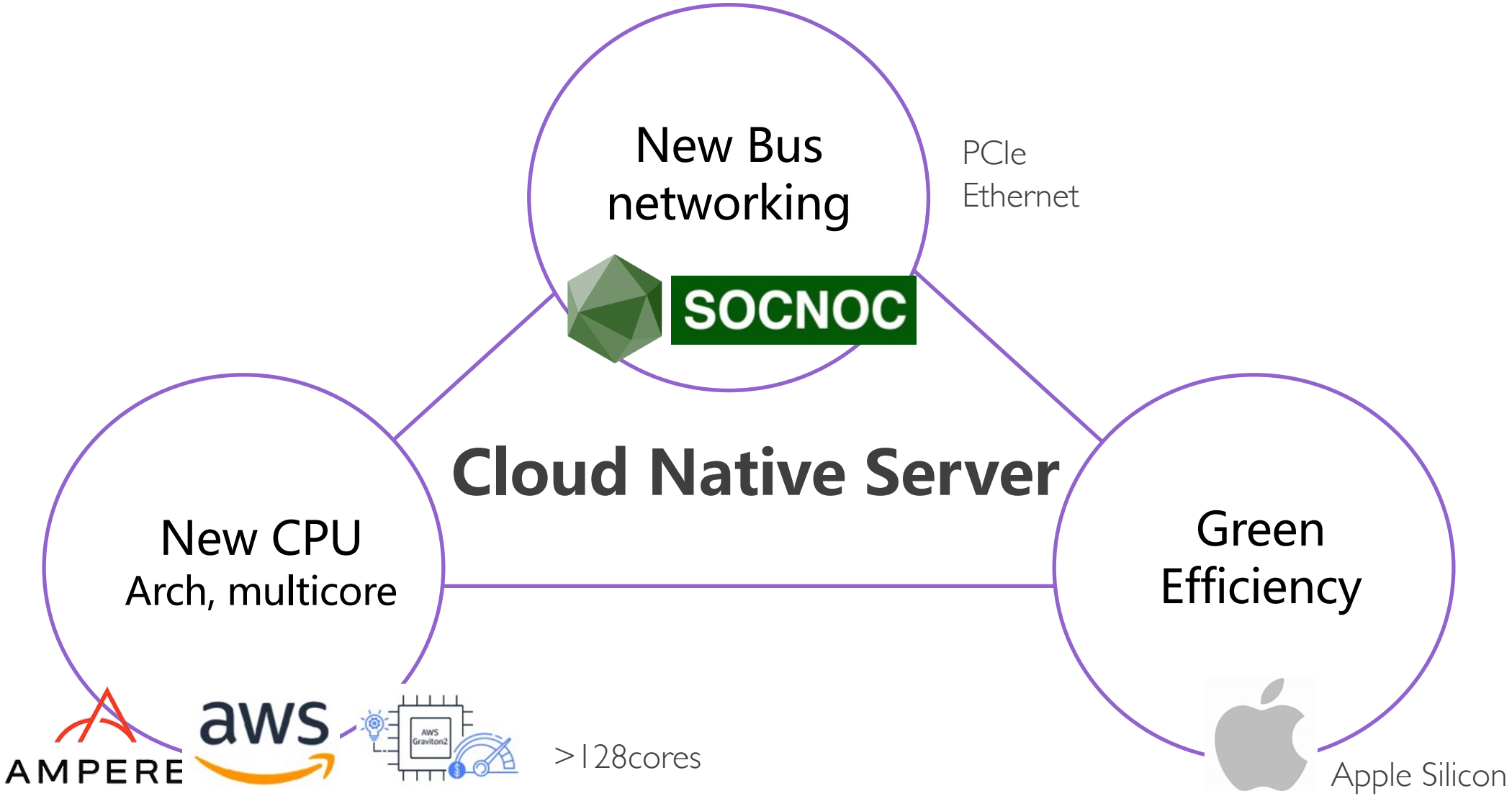
Consolidate network and link layer

	CPU/SoC #	PCIe Ports	Ethernet Switch Ports	Bottleneck
Elephant (HPC server)	2~4	24~36	2~4	Ethernet / CPU
Ants (microserver)	8~24	2-4	8~24	Ethernet Latency

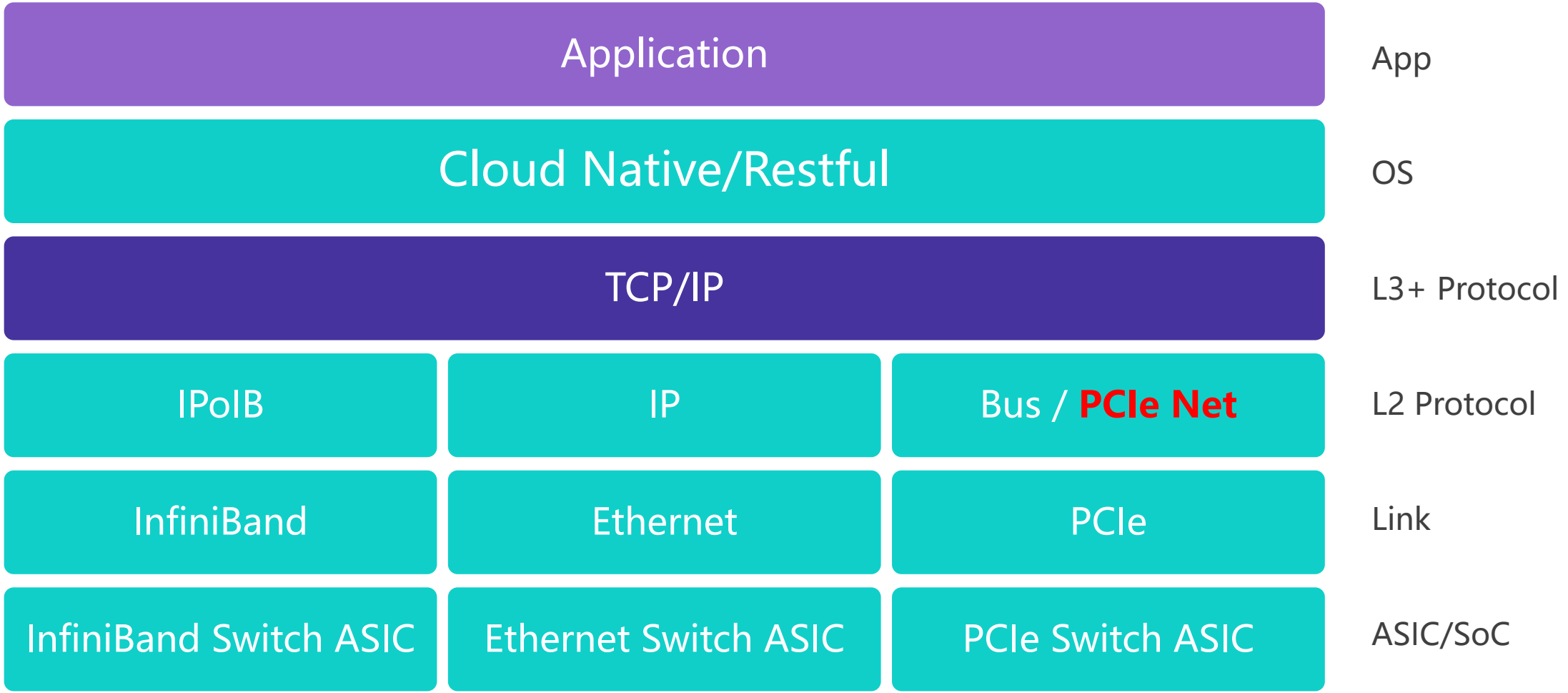
Solution: Merge PCIe Bus and Ethernet Net

PCIe Net

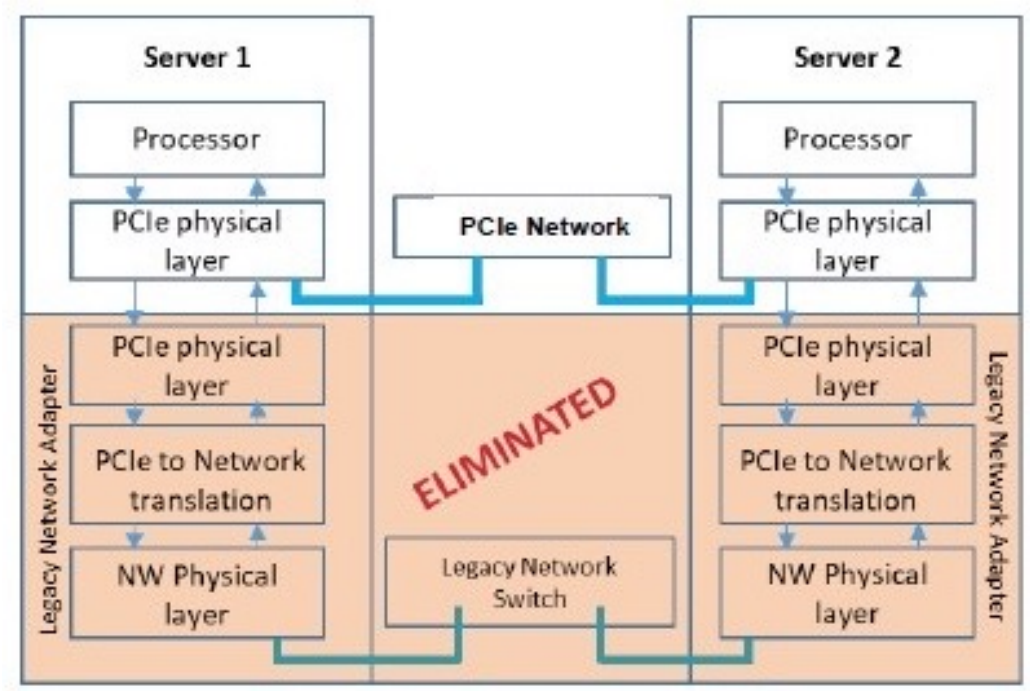
Next-generation System Design

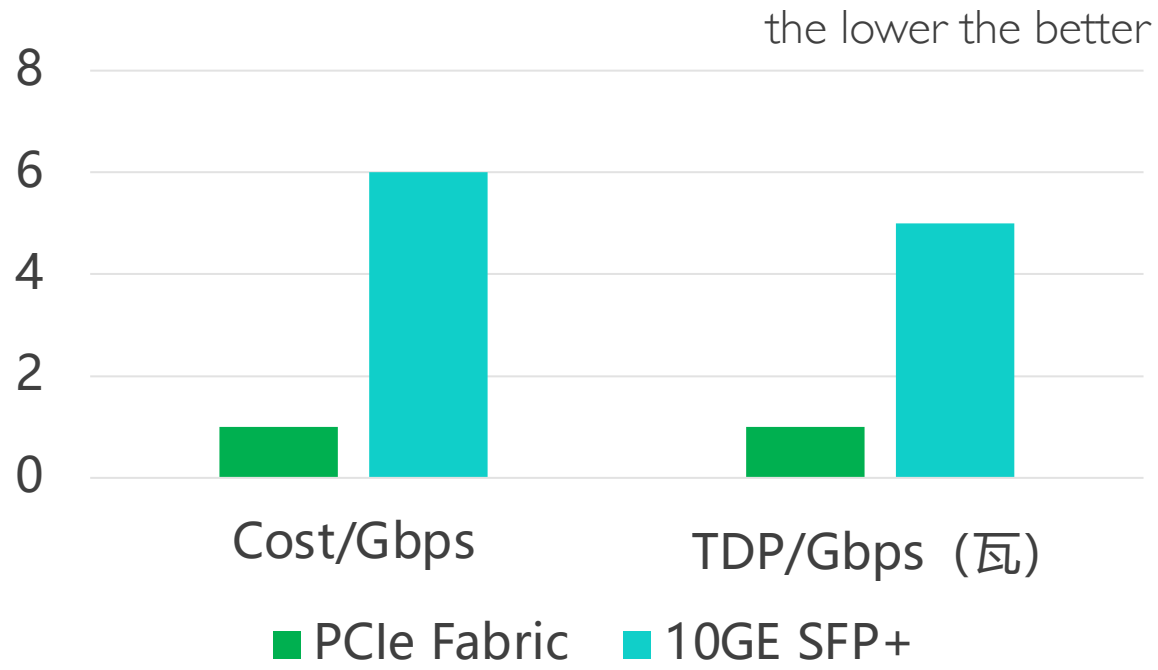


Networking Technologies for Clusters

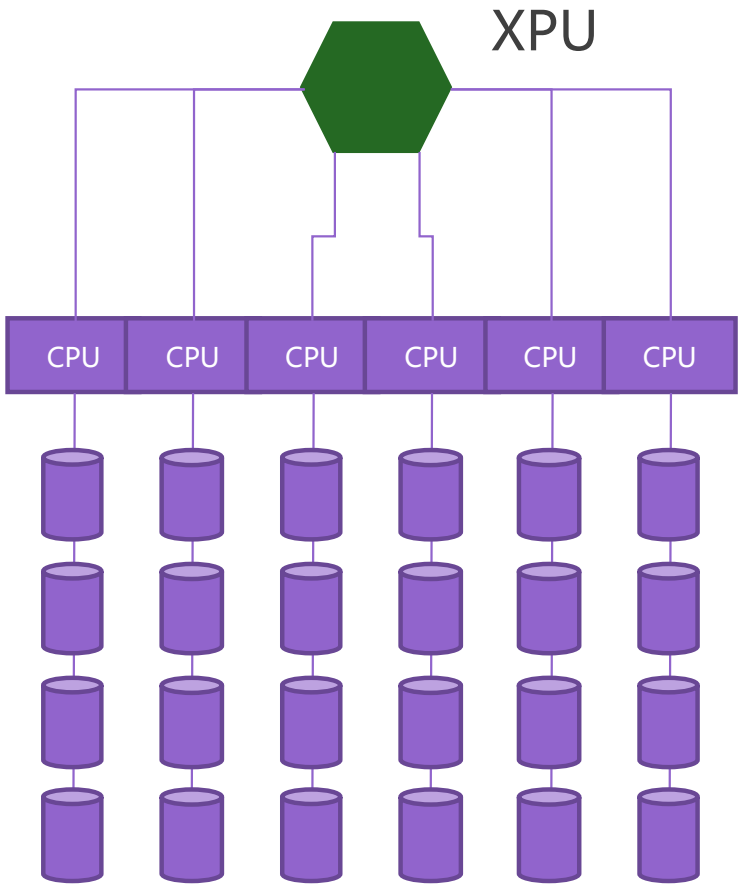


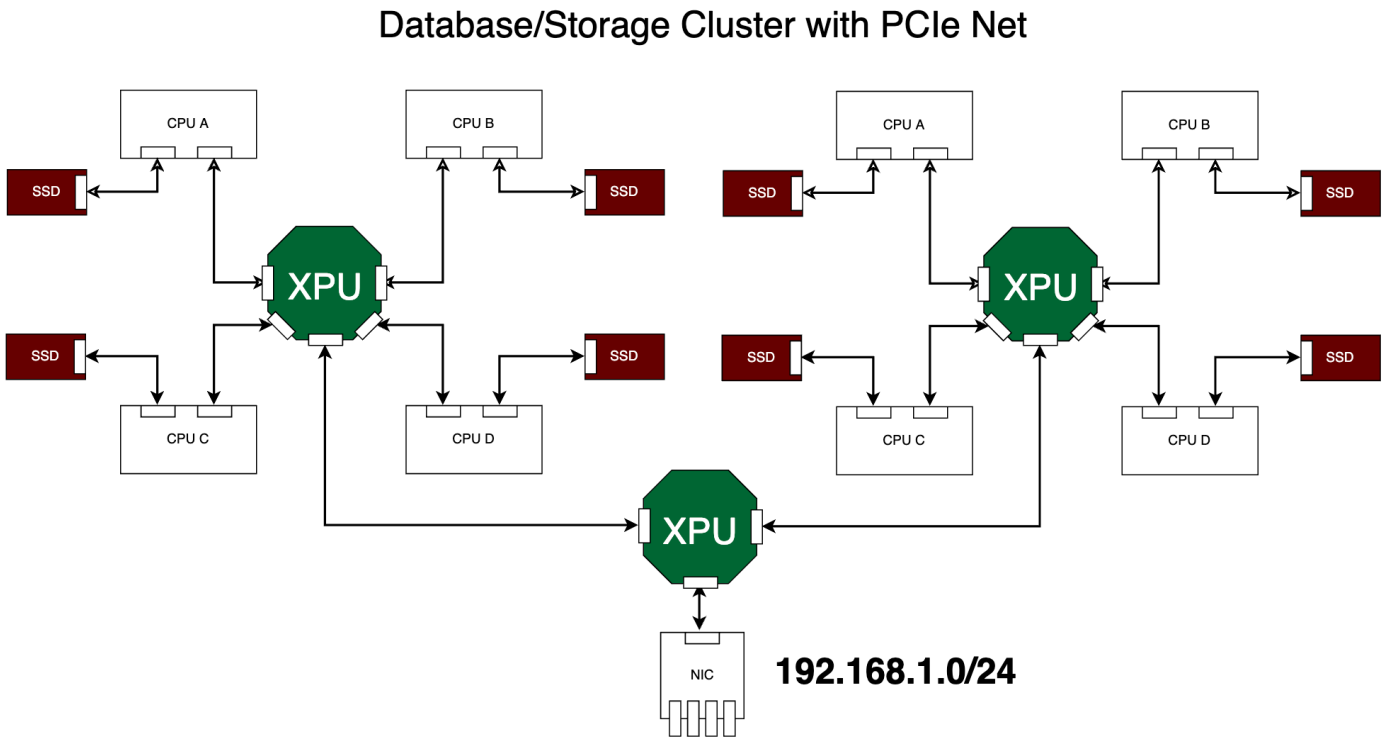
Extending PCIe Transport with Virtual NIC





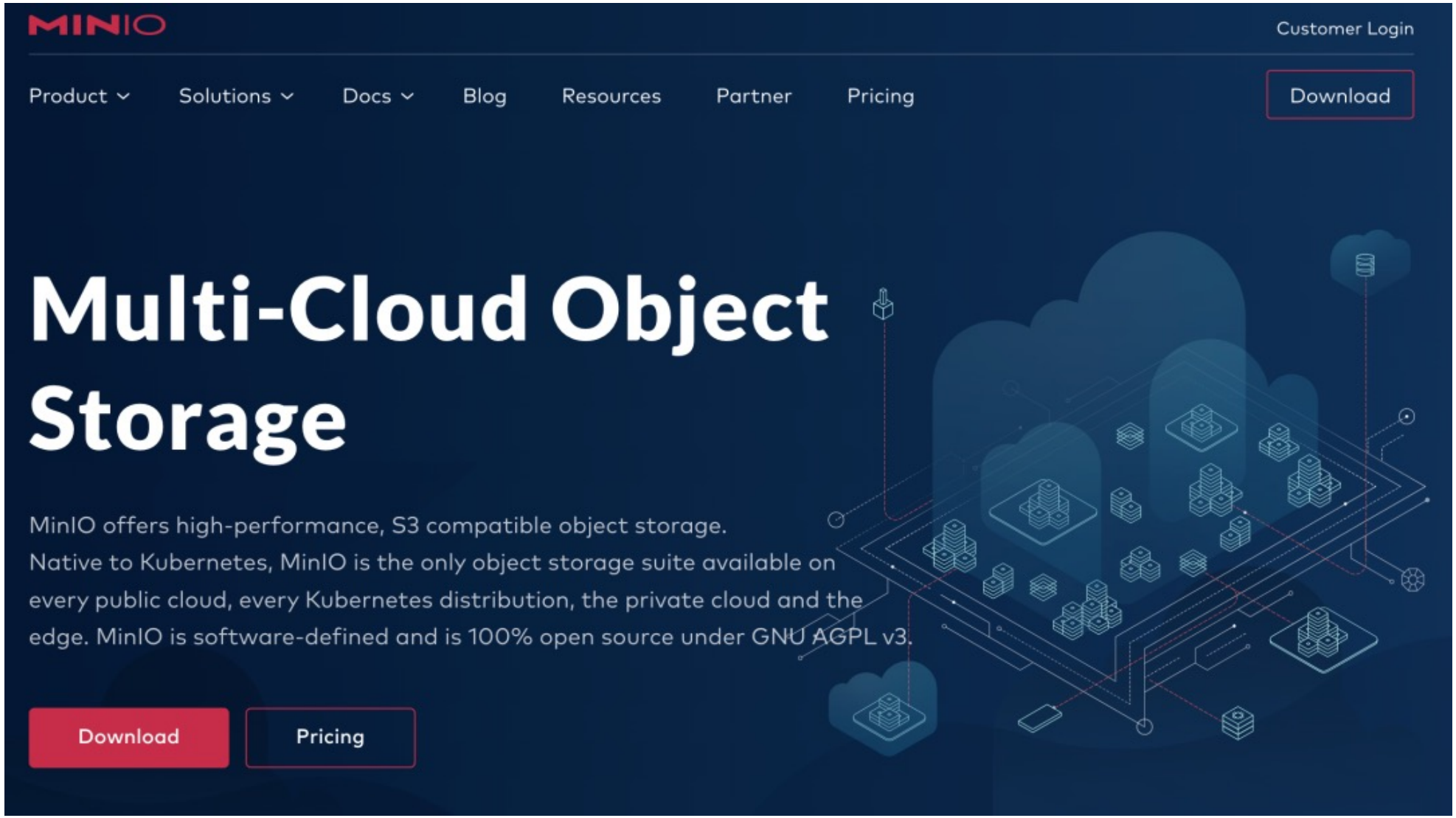
PCIe Net Cluster: Continuous Memory Space





each CPU "share" the same memory space, however communicating with TCP/IP

Demo: Object Storage MinIO over PCIe Net



Testbed: Four nodes (16 SSD)

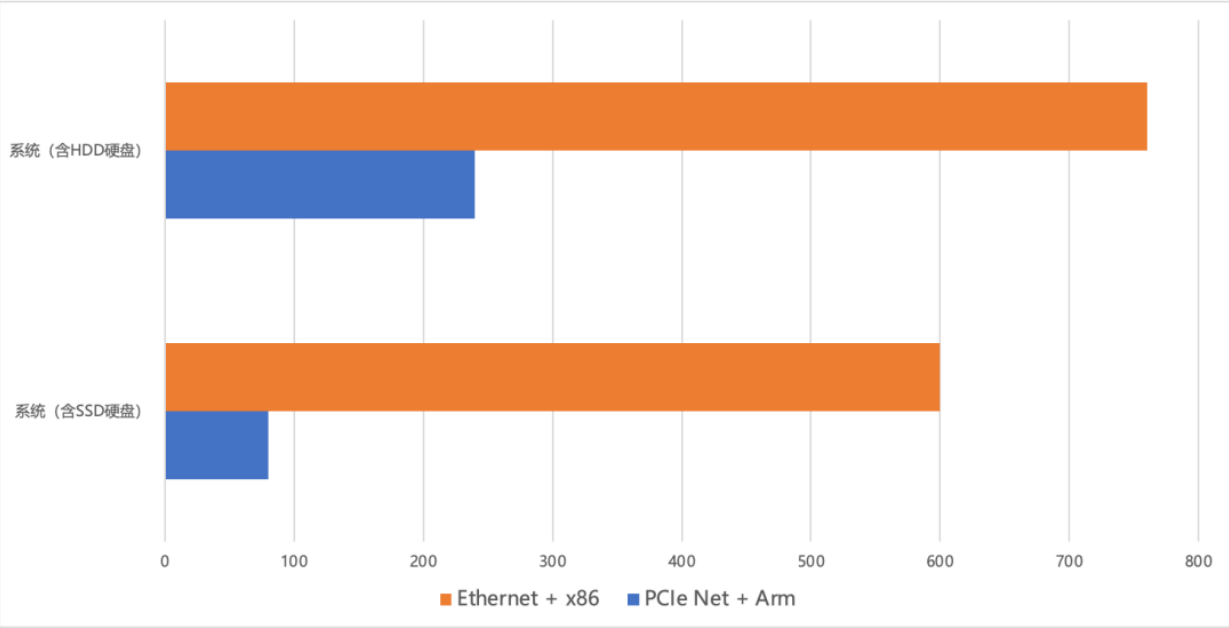
	#node	CPU	MEM	Storage	网络
VPC（x86）	4	24x2.5GHz	48GB	4x250GB	10GbE
IEC（arm）	4	24x1.0GHz	32GB	4x250GB	PCIe Net

	VPC（x86）	IEC（arm）
os	ubuntu 20.04	ubuntu 20.04 (arm)
minIO	RELEASE.2022-07-30	RELEASE.2022-07-30

```
minio server http://min{1...4}.pcie.net:9000/mnt/disk{1...4}/minio
```

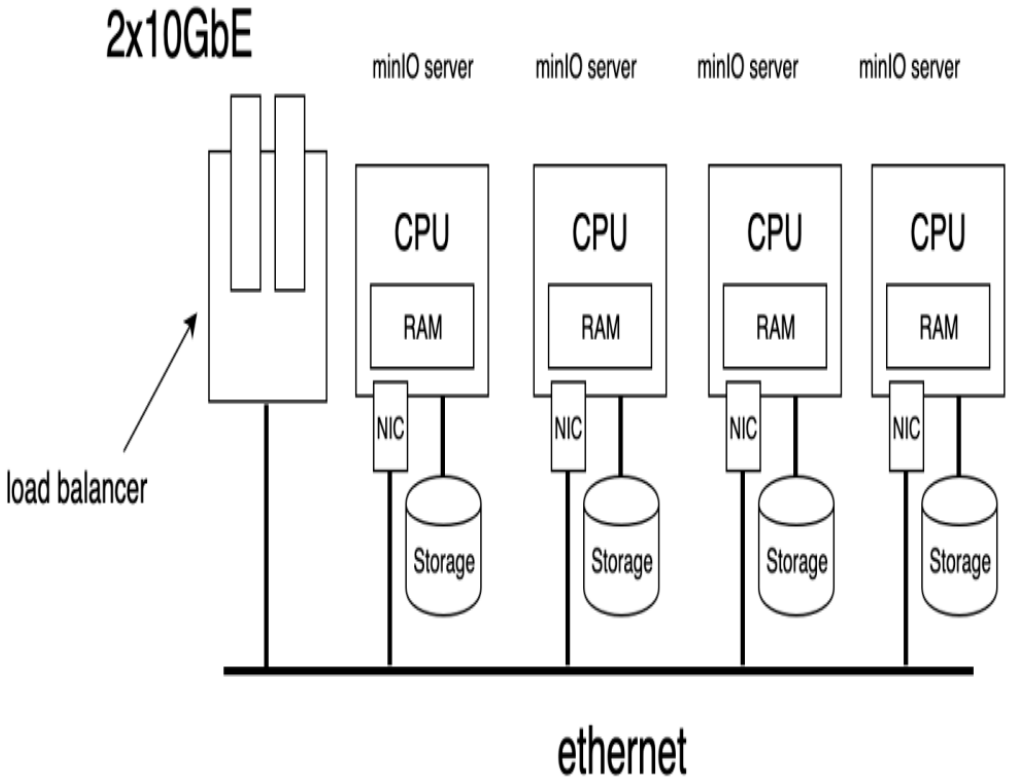
Power

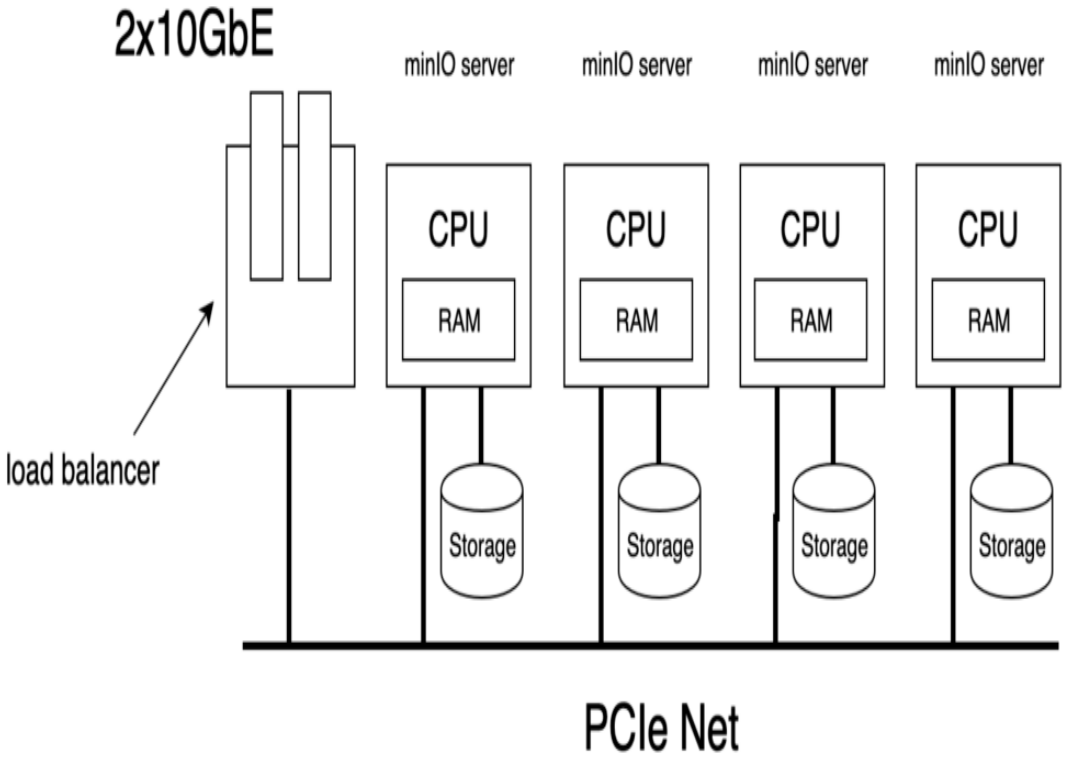
功耗（瓦）	系统（含SSD硬盘）	系统（含HDD硬盘）
PCIe Net + Arm	80	240
Ethernet + x86	600	760



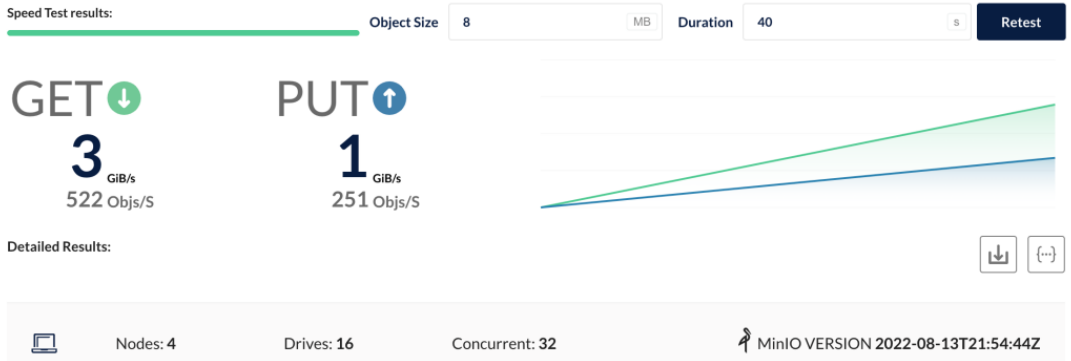
PCIe Net + arm CPU save lots of energy

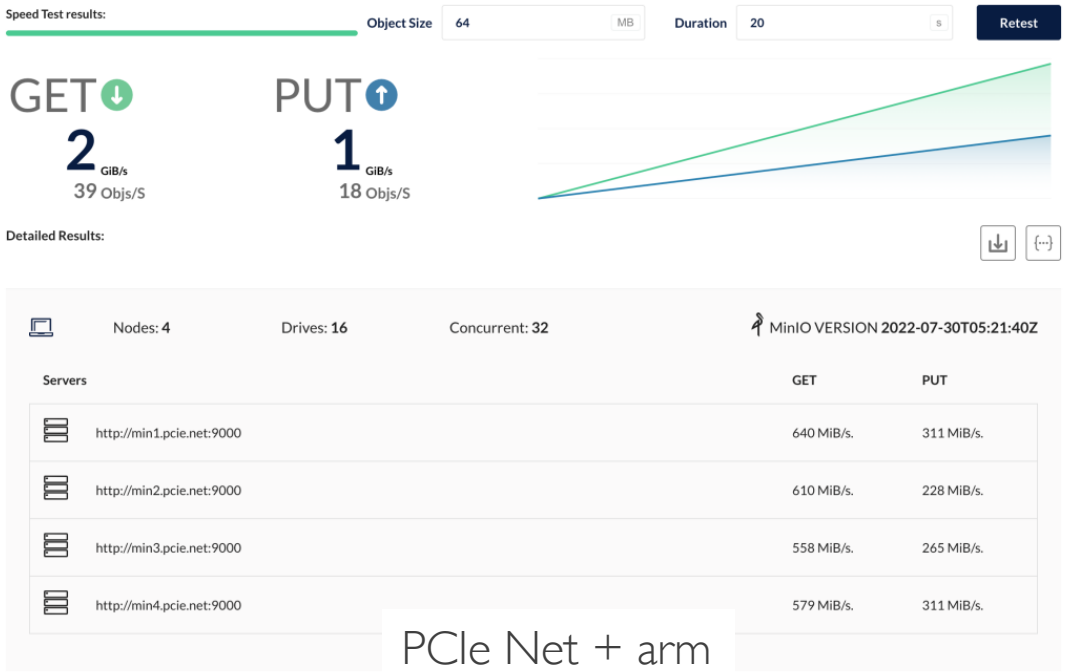
Network


2x10GbE
minIO server
minIO server
minIO server
minIO server
CPU
CPU
CPU
CPU
RAM
RAM
RAM
RAM
NIC
NIC
NIC
NIC
Storage
Storage
Storage
Storage
load balancer
ethernet

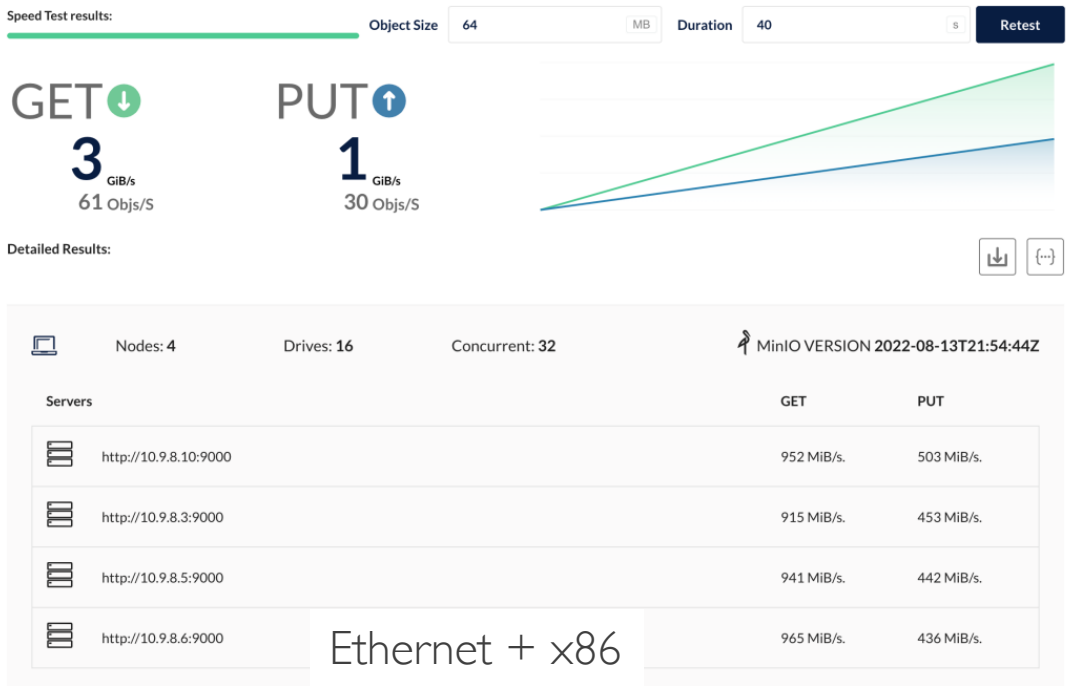

2x10GbE
minIO server
minIO server
minIO server
minIO server
CPU
CPU
CPU
CPU
RAM
RAM
RAM
RAM
Storage
Storage
Storage
Storage
load balancer
PCIe Net

Speed Test





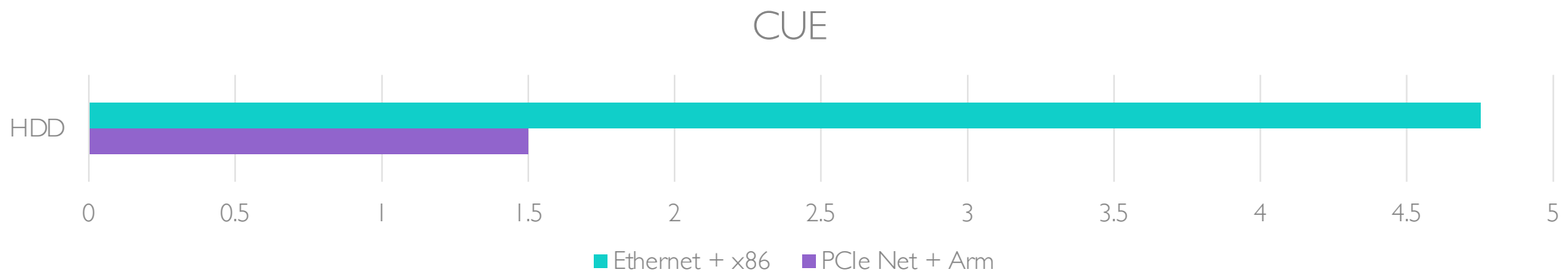
PCIe Net + arm

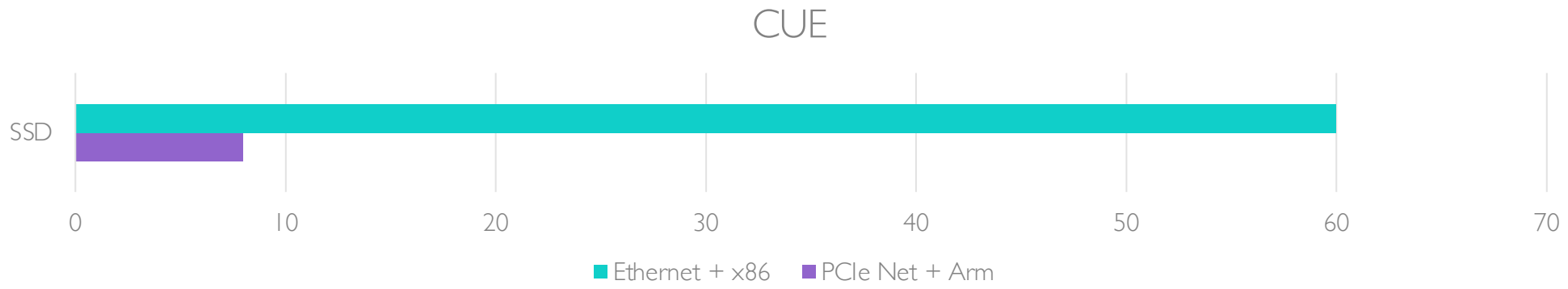


Ethernet + x86

Carbon Usage Effectiveness (CUE)

$$CUE = \frac{Hard\ Drive\ TDP + Bare\ System\ TDP + Networking\ TDP}{Hard\ Drive\ TDP}$$



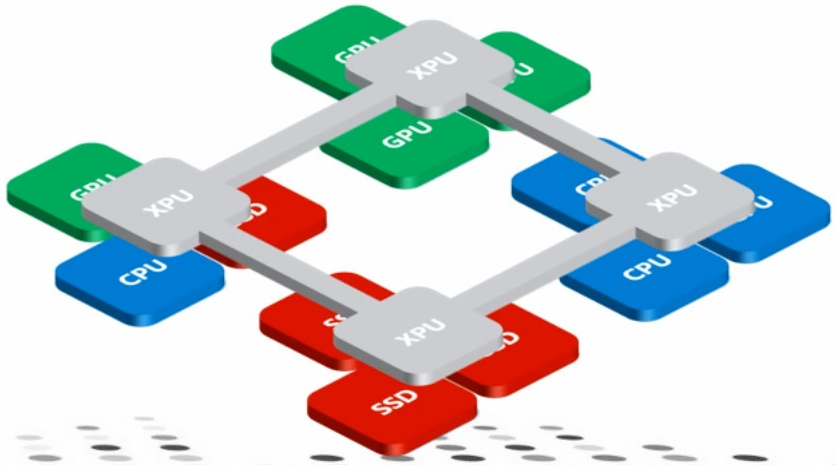


SOCNOC

SOCNOC Co., LTD

Distributed storage solution based on PCIe Net

All in PCIe: Just Connecting



SOCNOC Co., LTD

https://www.socnoc.ai

Thank you !

www.socnoc.ai