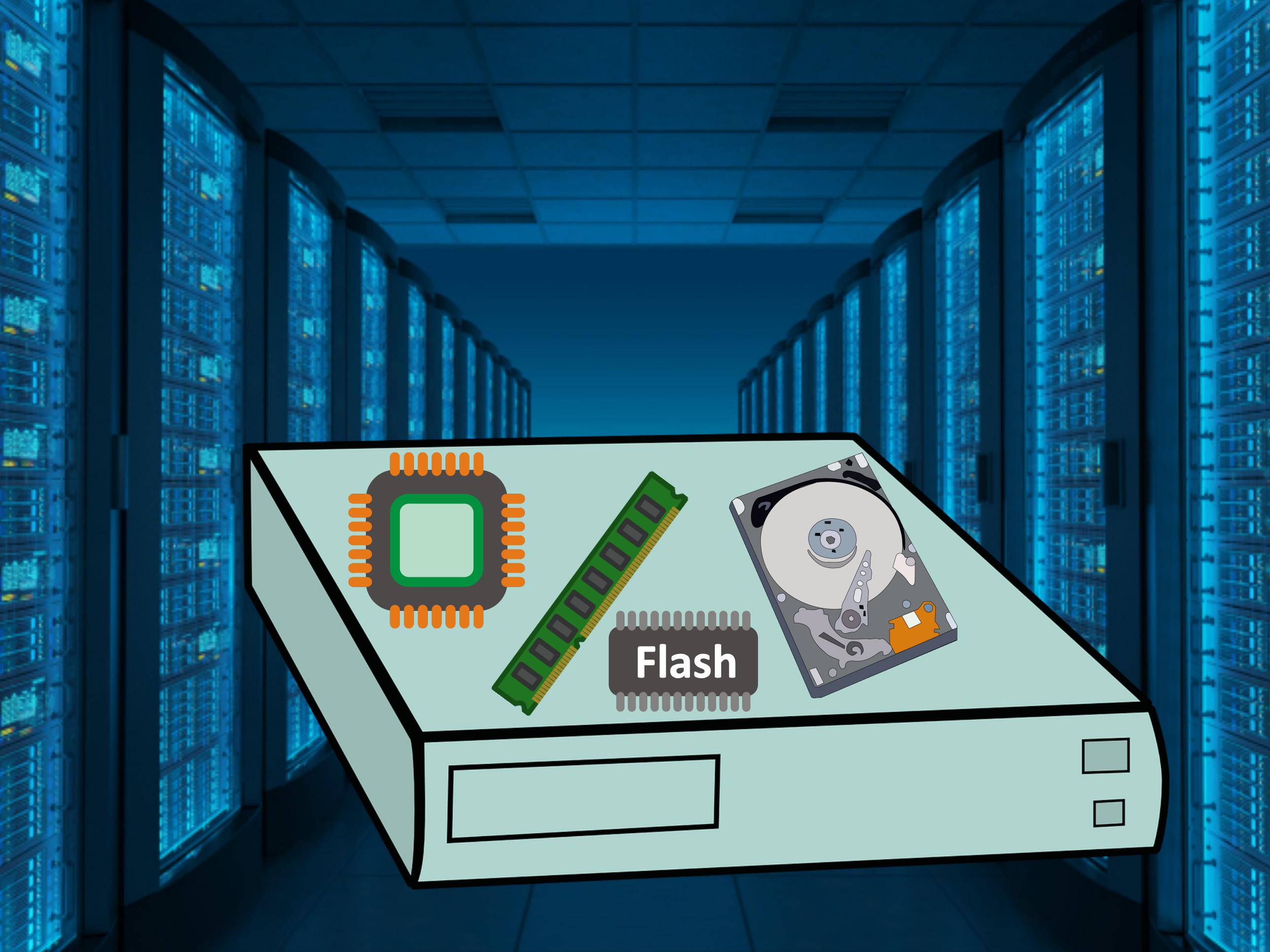
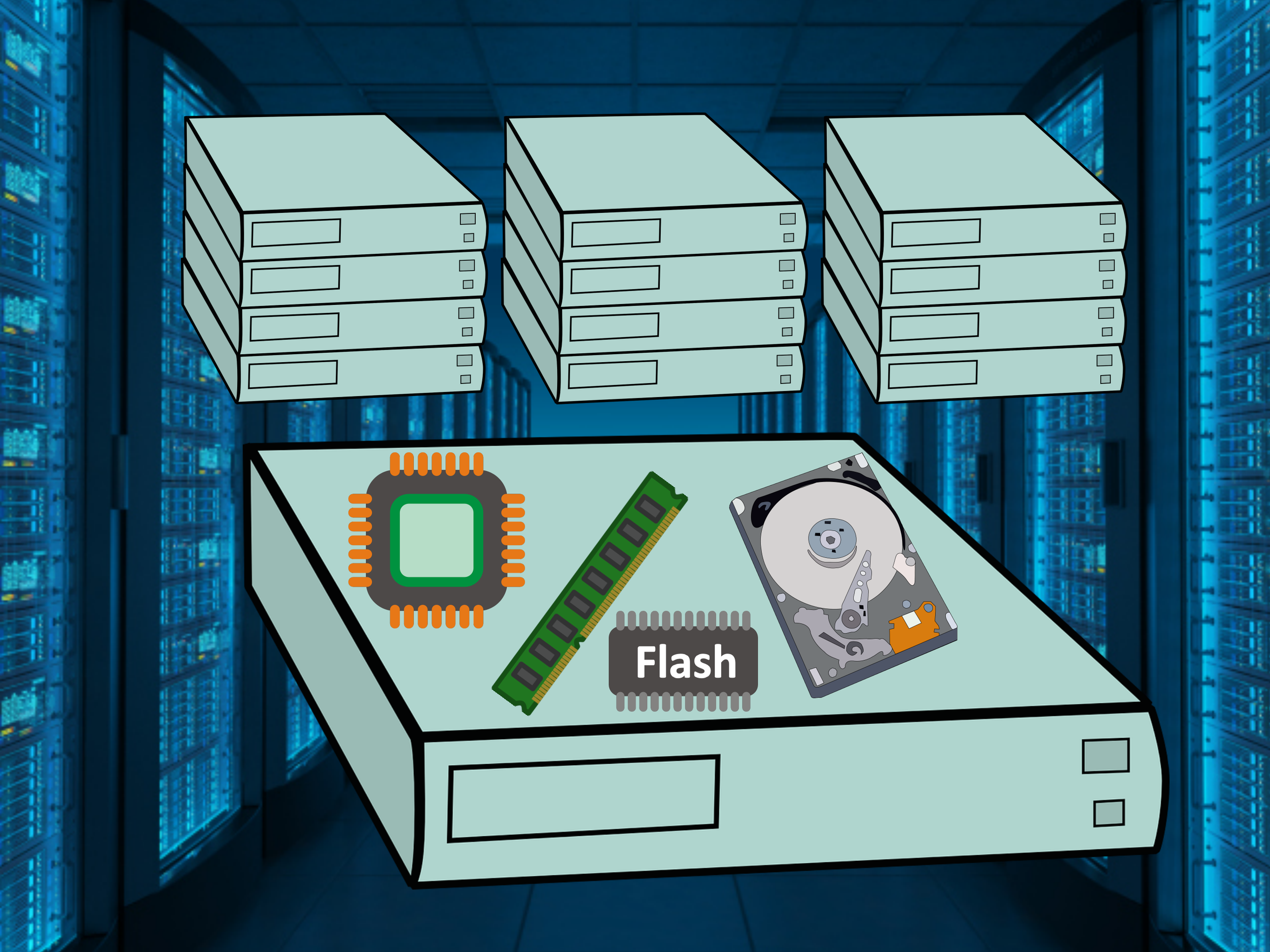
Disaggregated Operating System

Yiying Zhang

Yizhou Shan, Yilun Chen, Yutong Huang, Sumukh Hallymysore

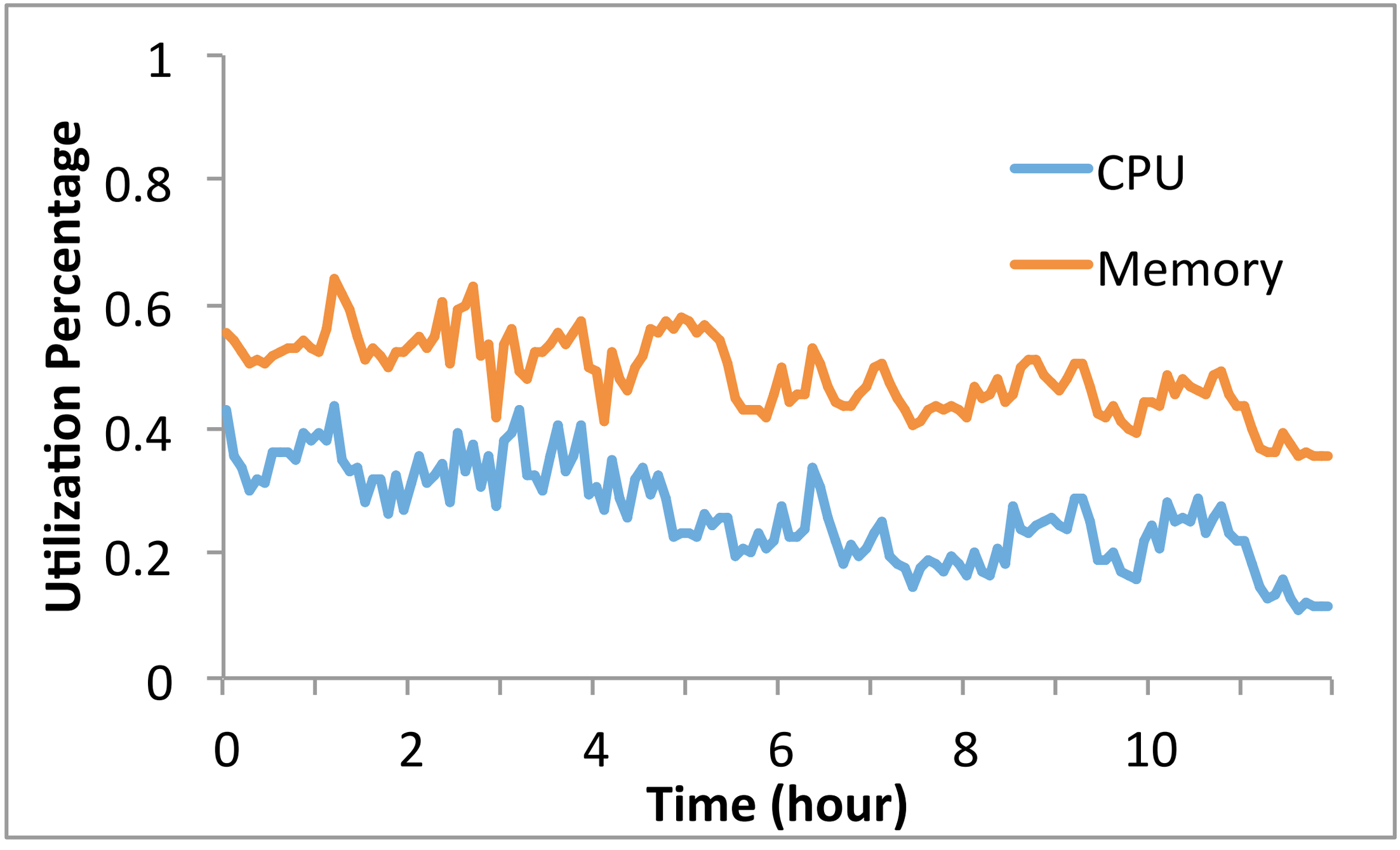

Flash


Flash

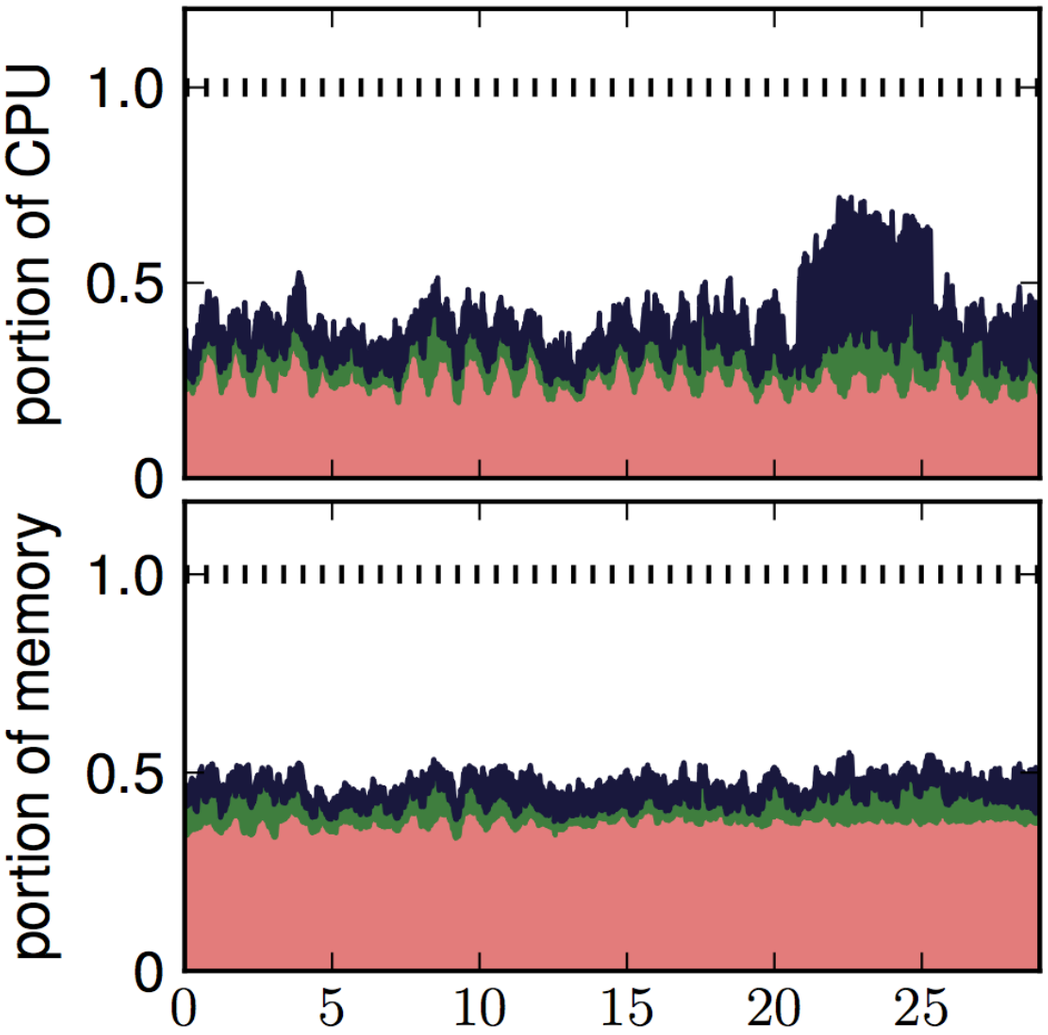
Monolithic Server

- Resource utilization
- Failure
- Flexibility

Alibaba Cluster Resource Utilization

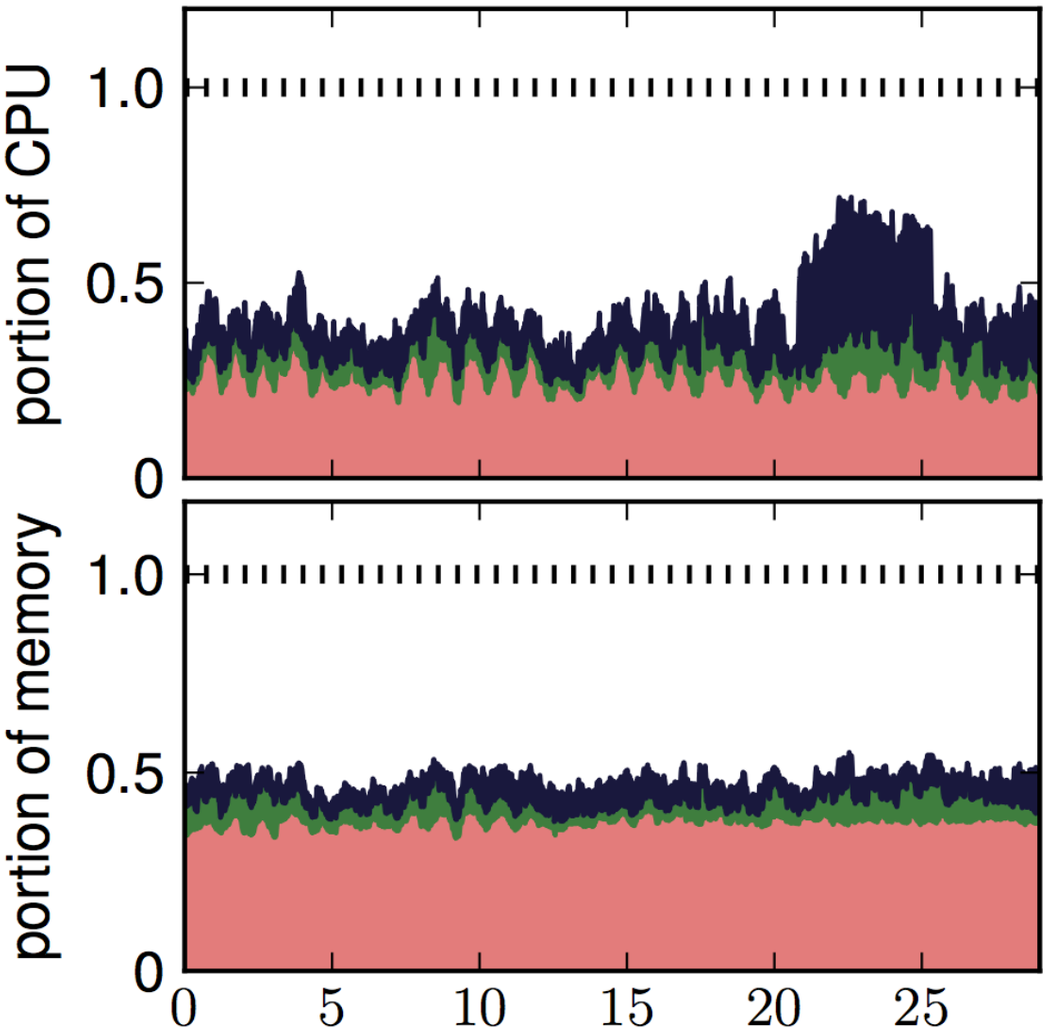


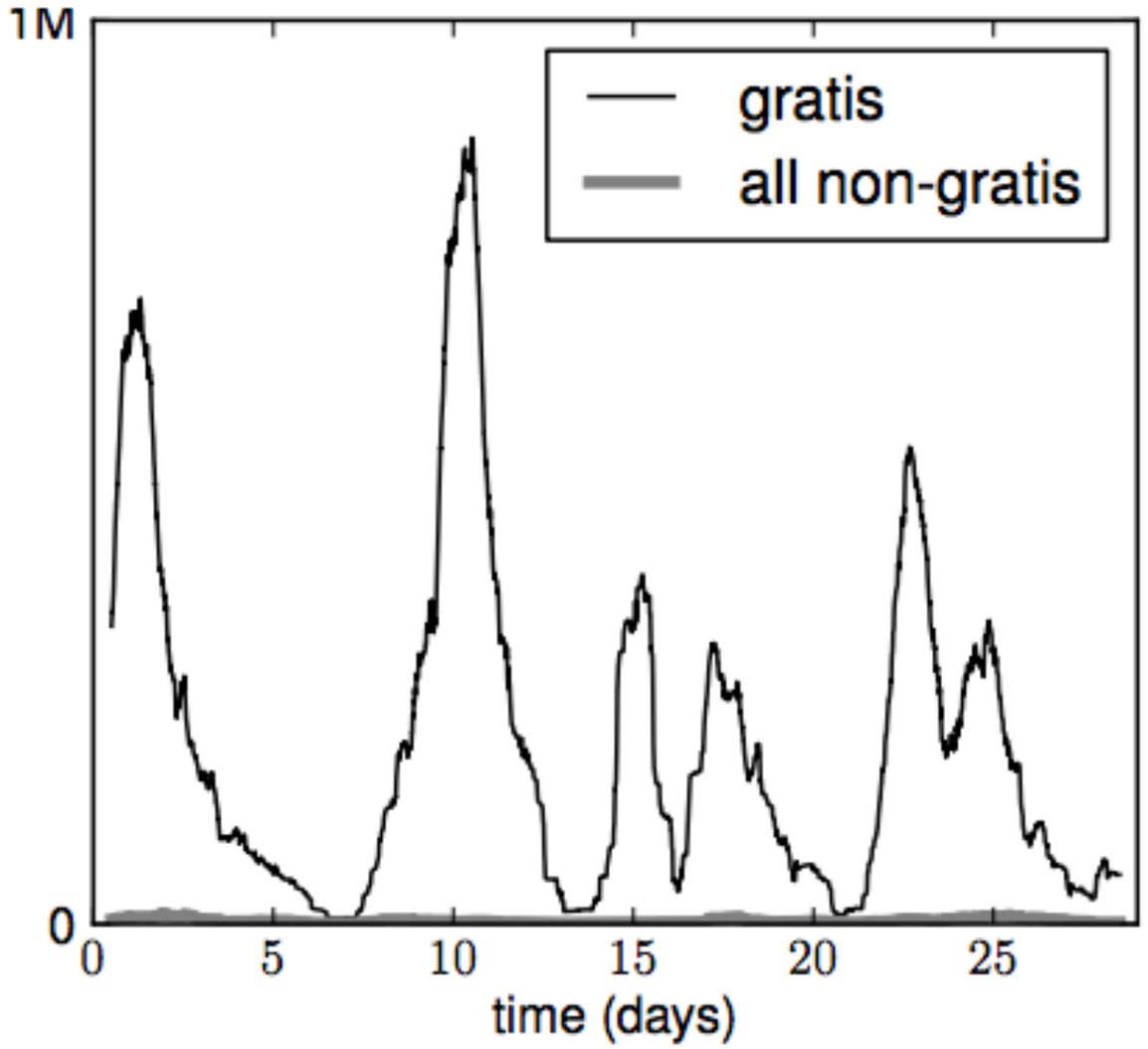
Google Cluster Resource Utilization



REISS, C. etal. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. SoCC'12

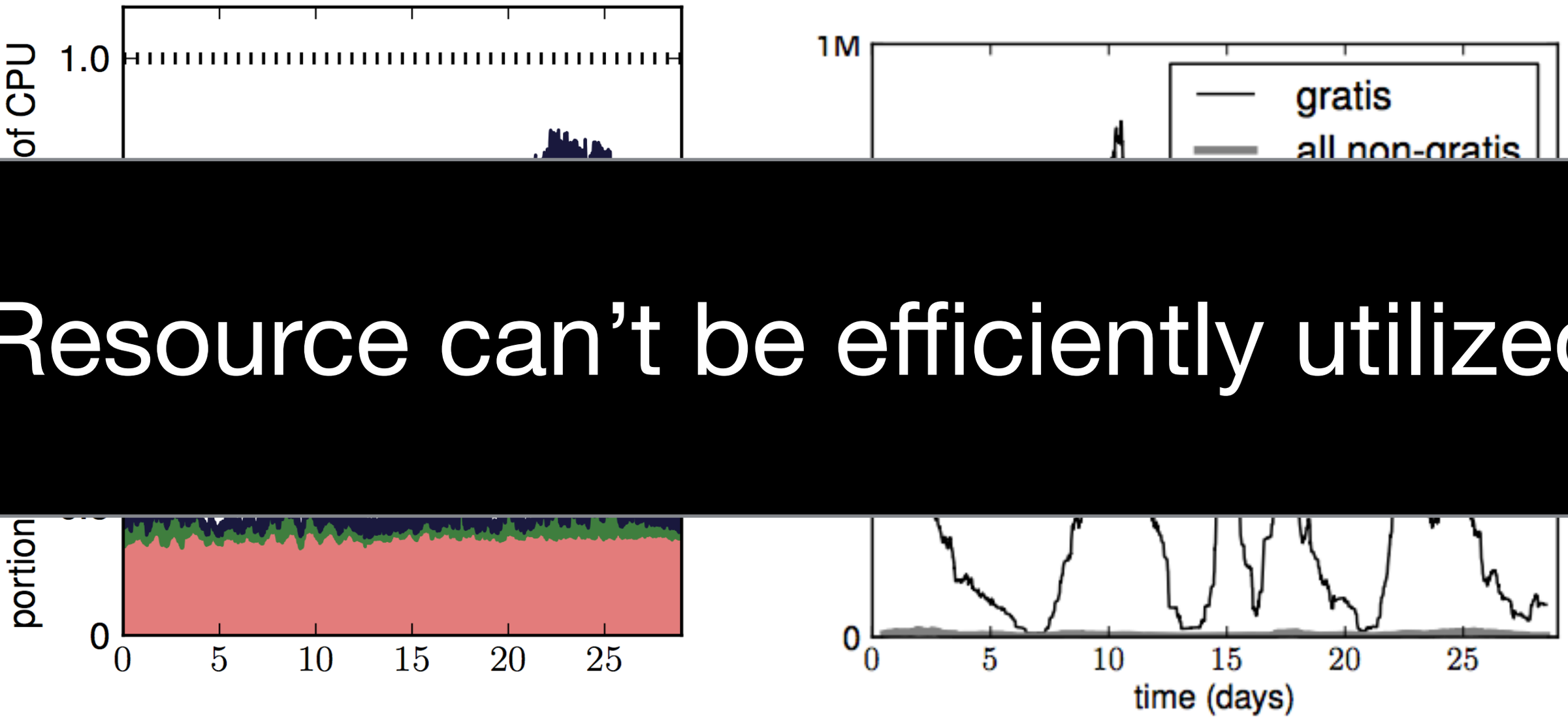
Google Cluster Resource Utilization





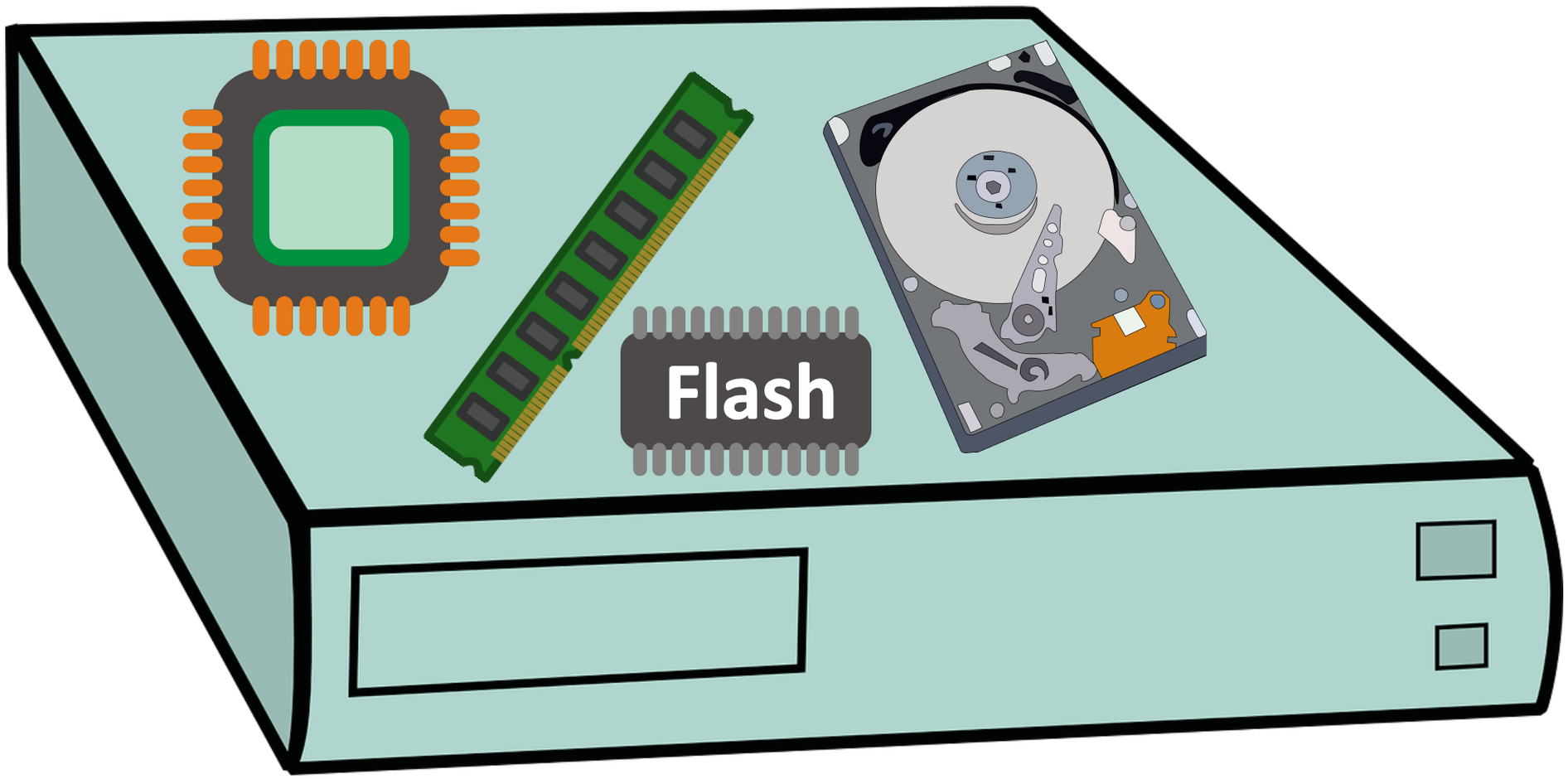
REISS, C. etal. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. SoCC'12

Google Cluster Resource Utilization



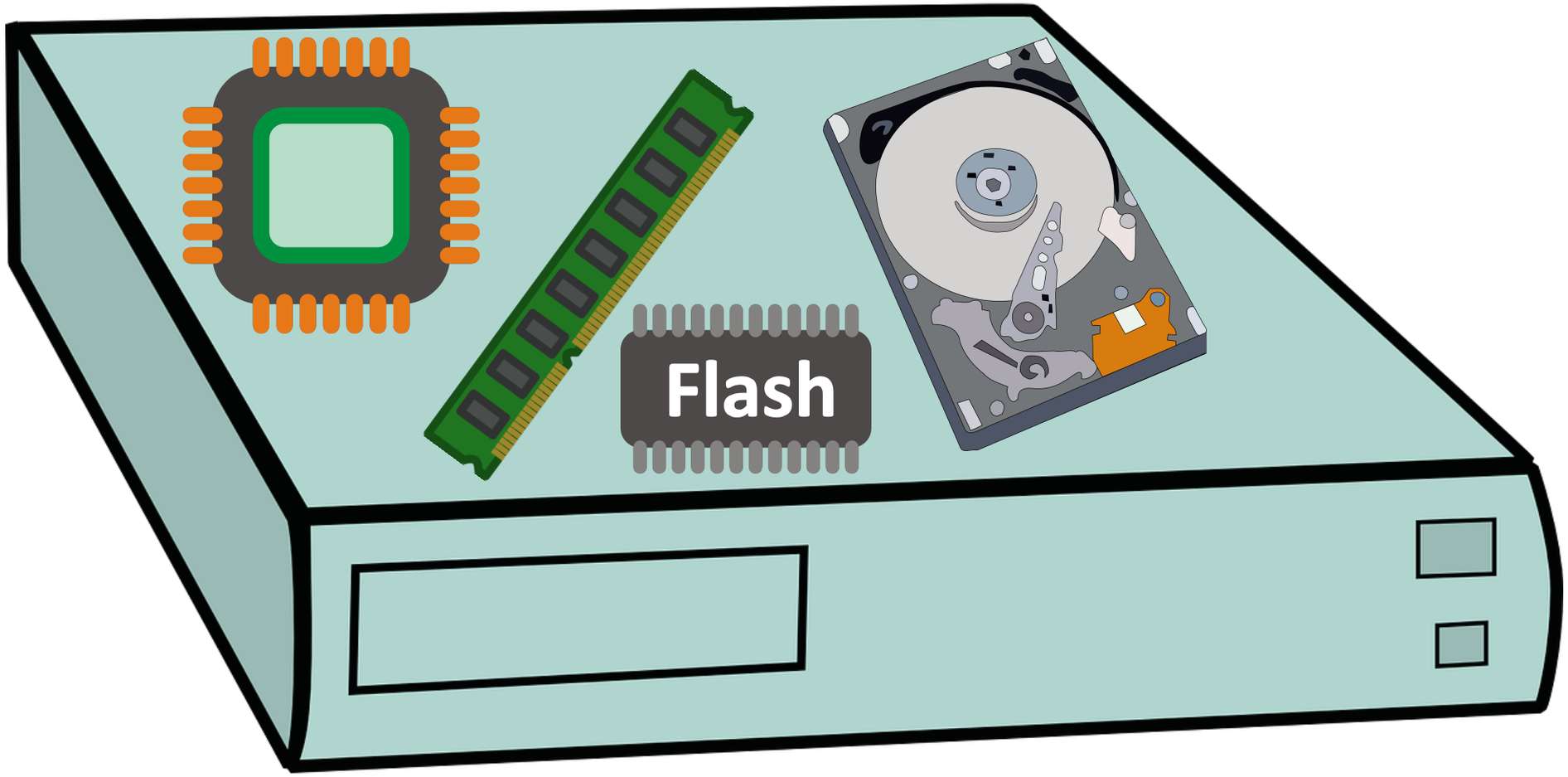
Resource can’t be efficiently utilized

REISS, C. etal. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. SoCC’12


Flash

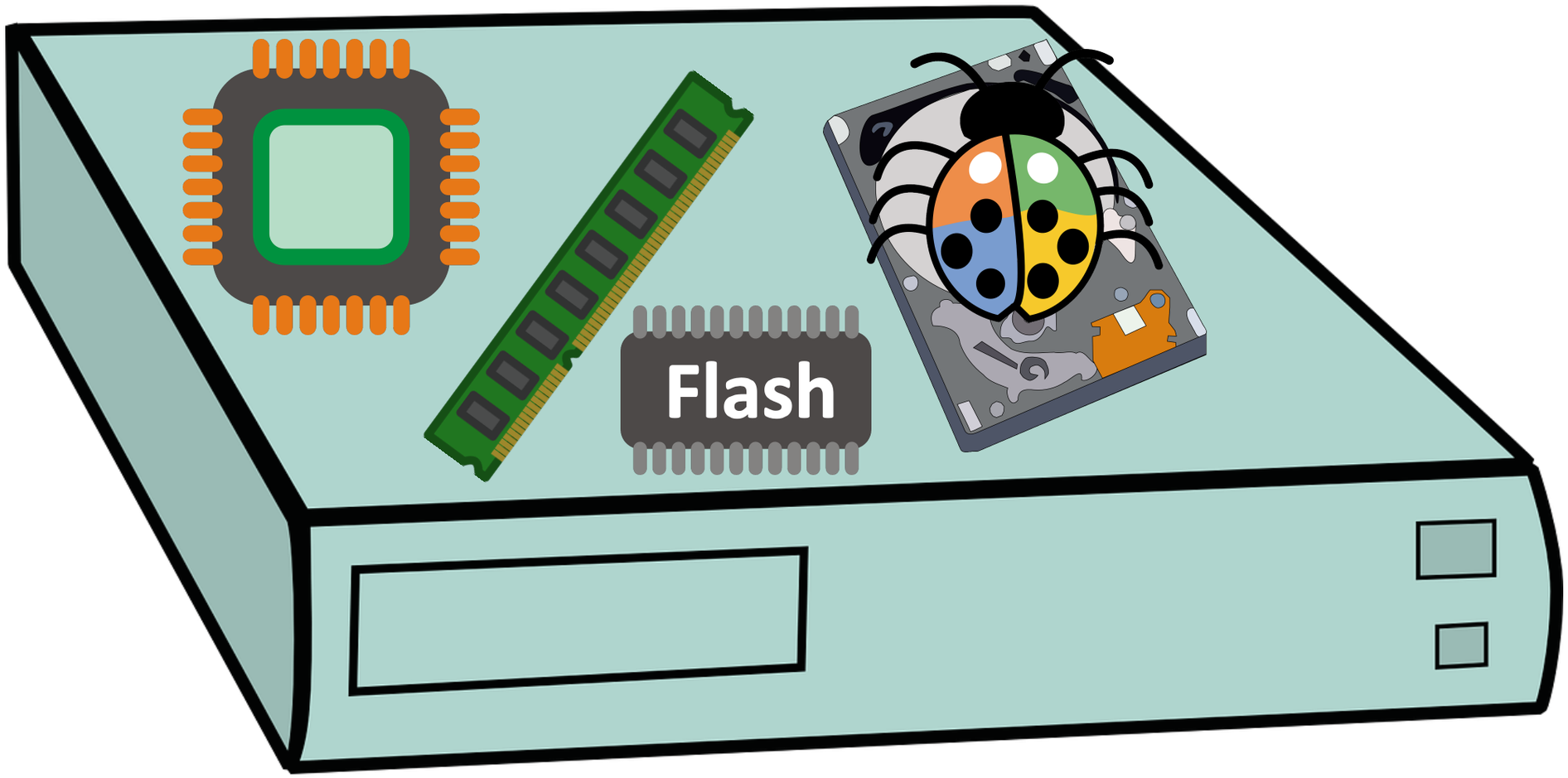
App App App App

OS/Hypervisor



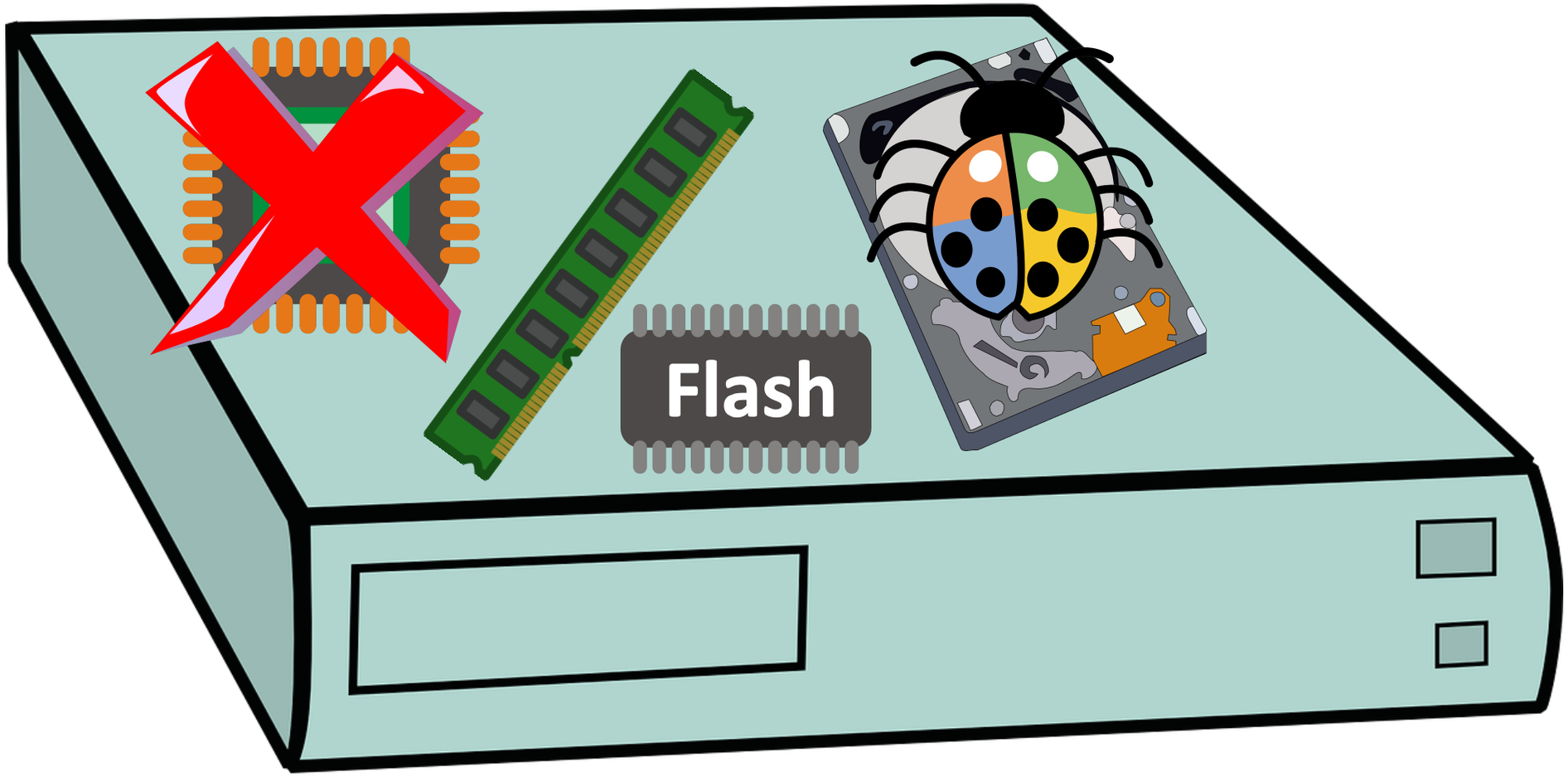
App App App App

OS/Hypervisor



App App App App

OS/Hypervisor

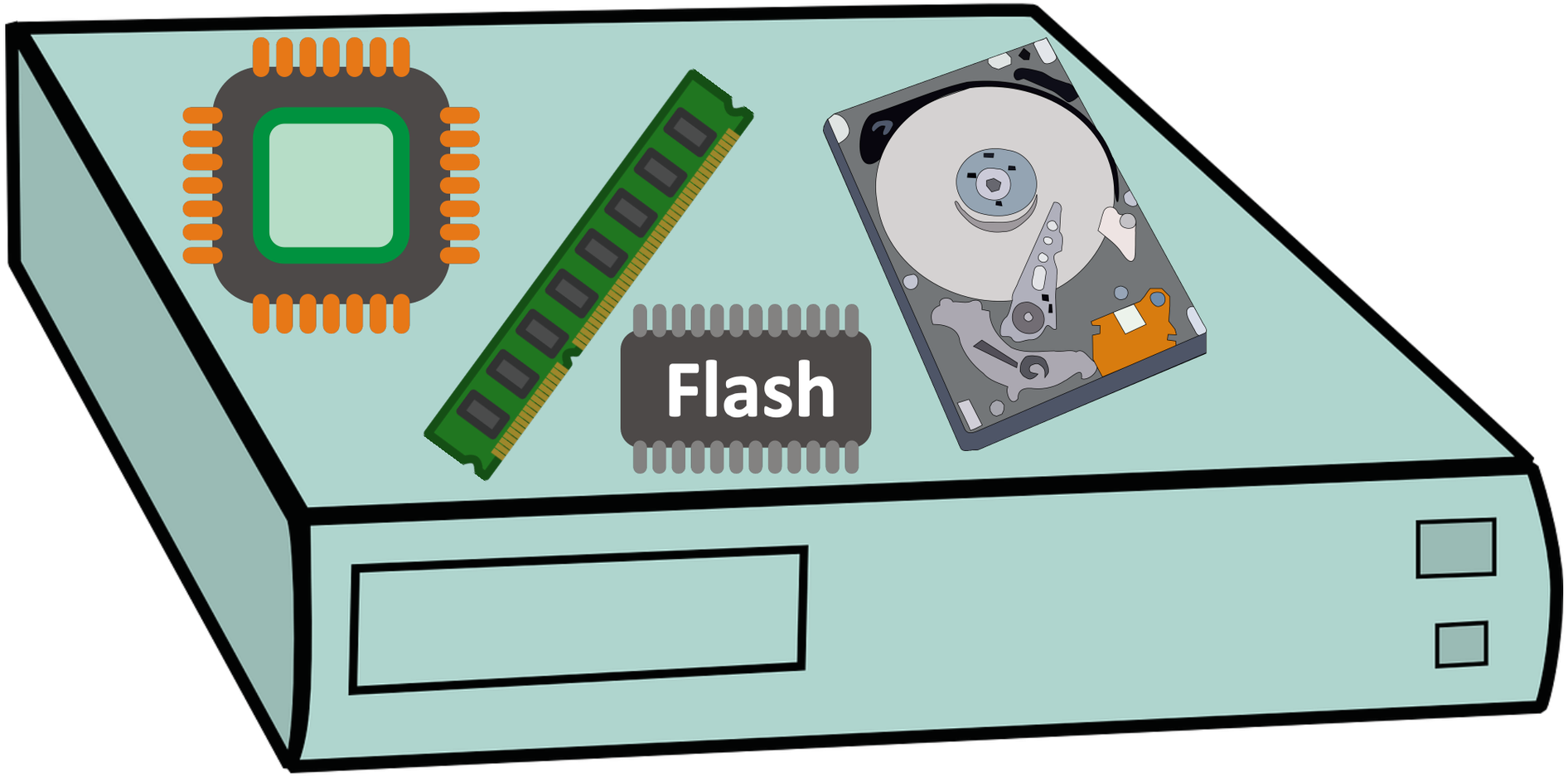



App
BOOM!!

No fine-grained failure handling

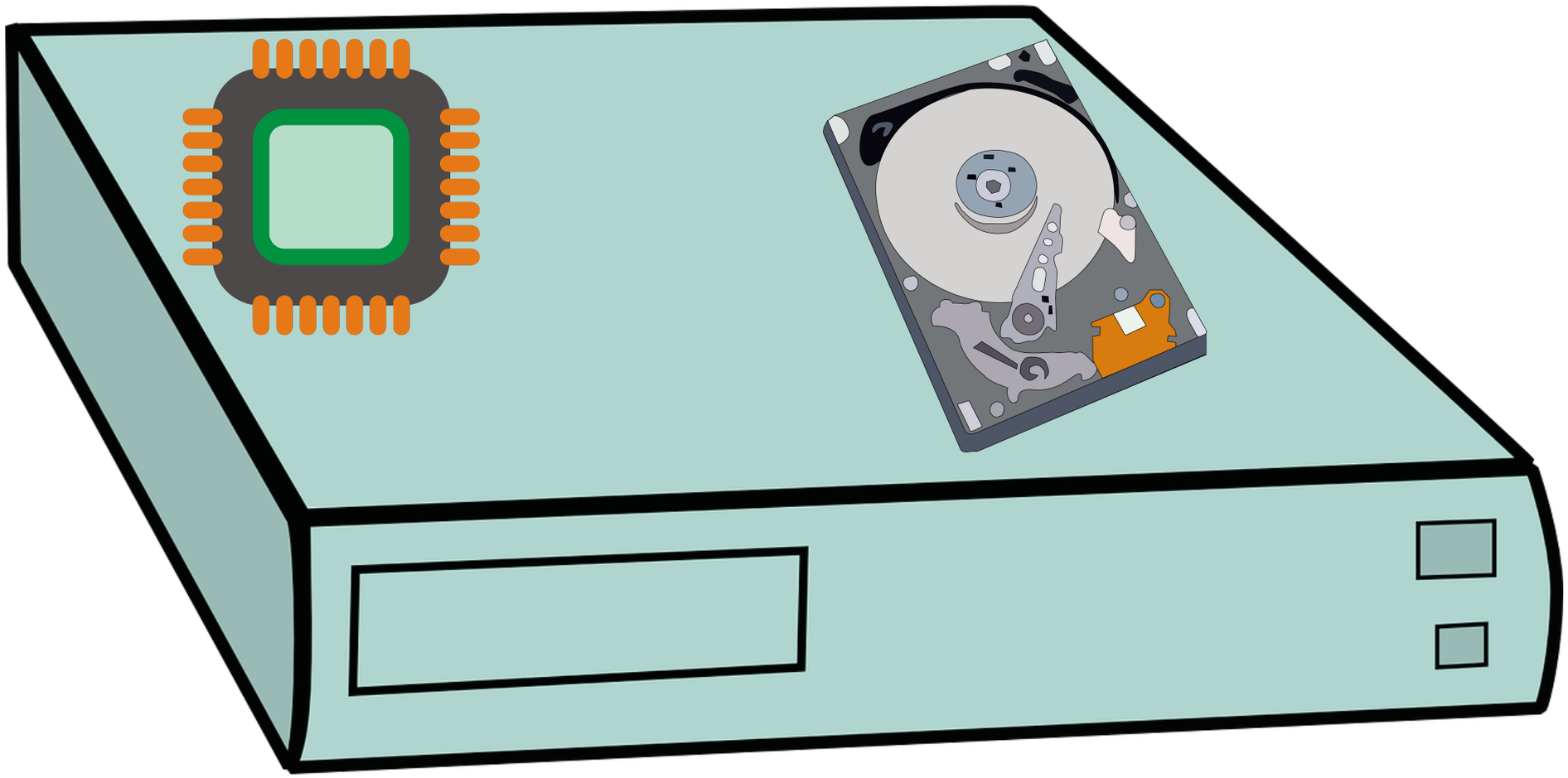
App App App App

OS/Hypervisor



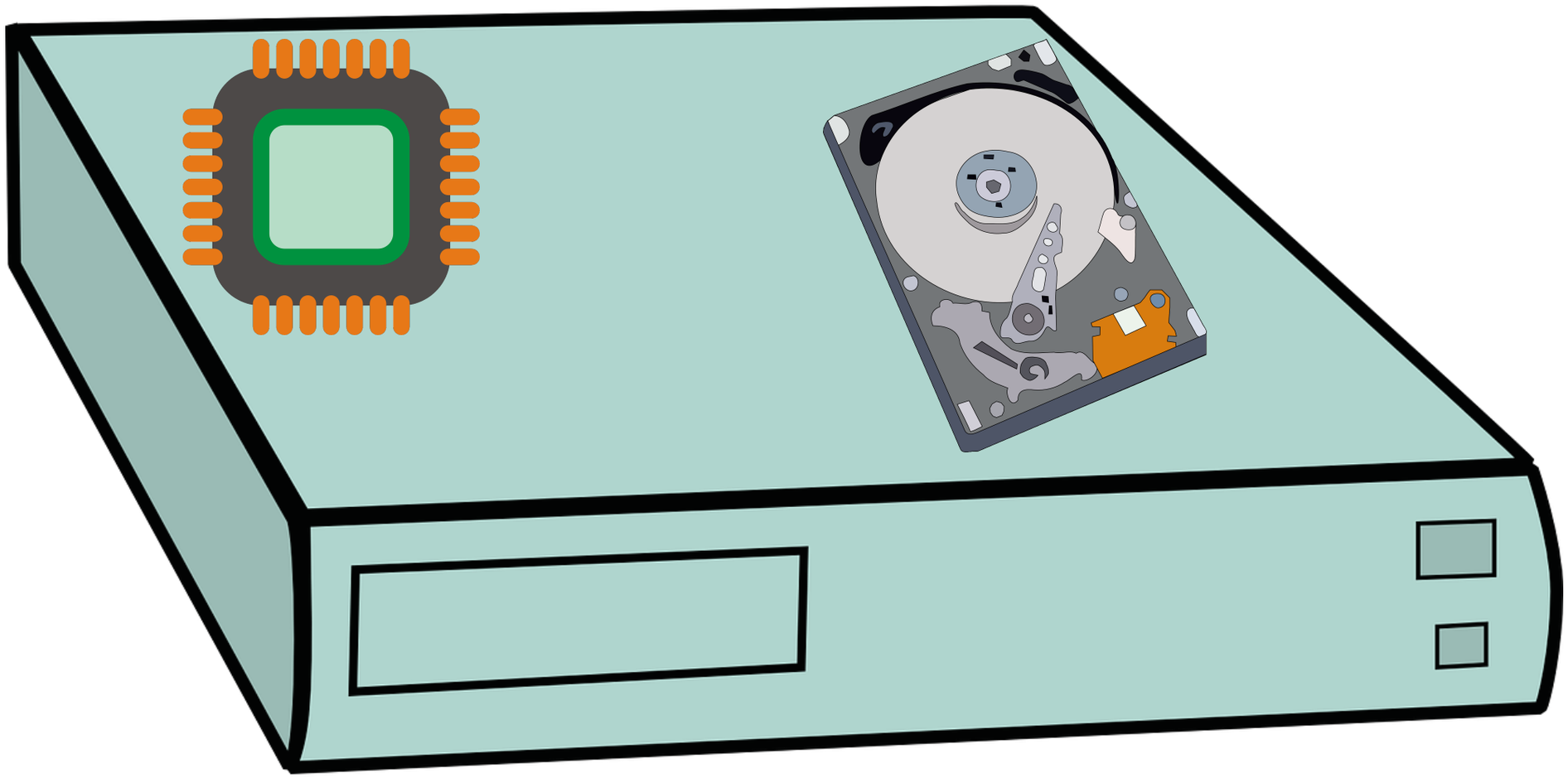
App App App App

OS/Hypervisor



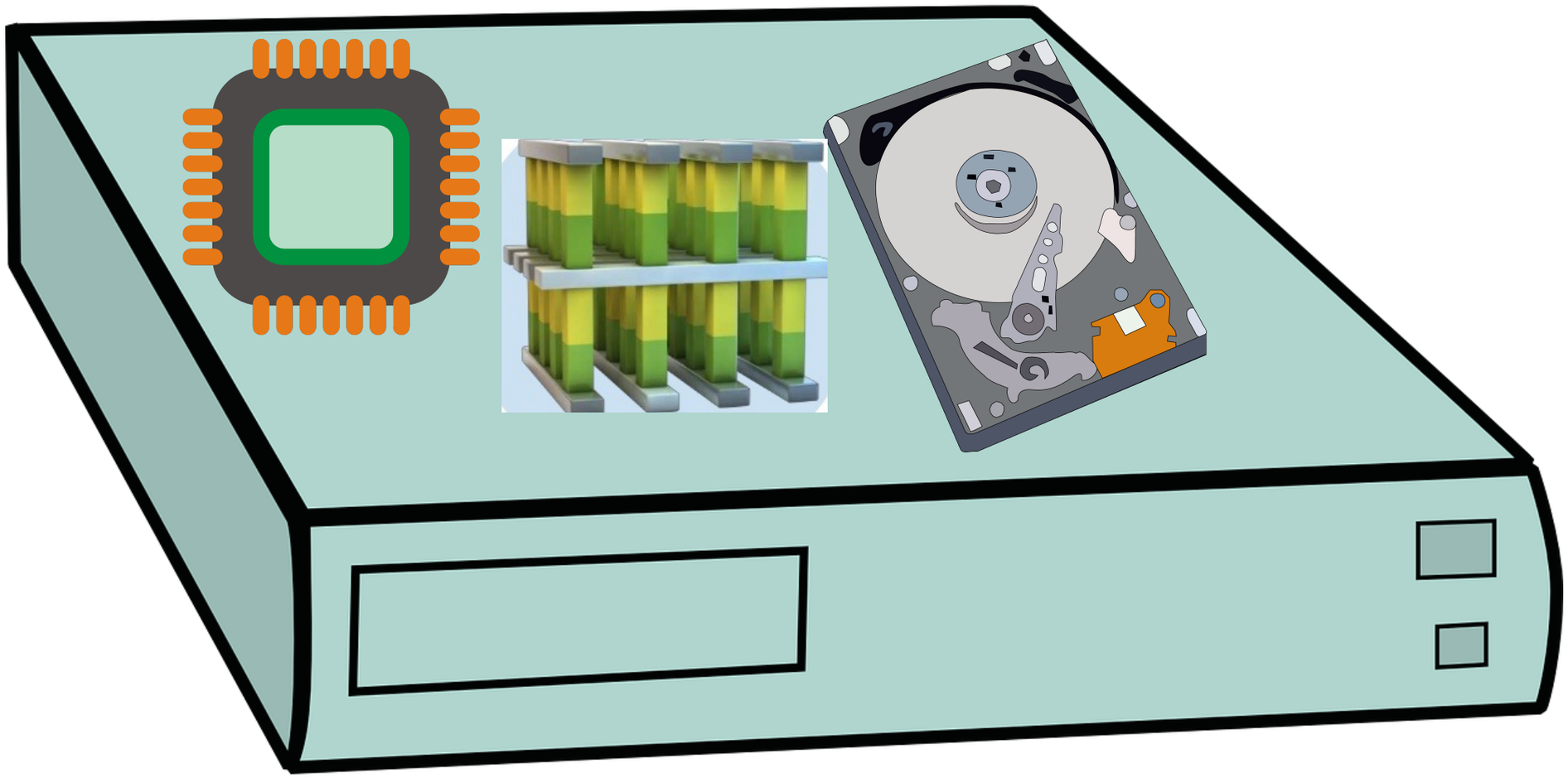
App App App App

OS/Hypervisor



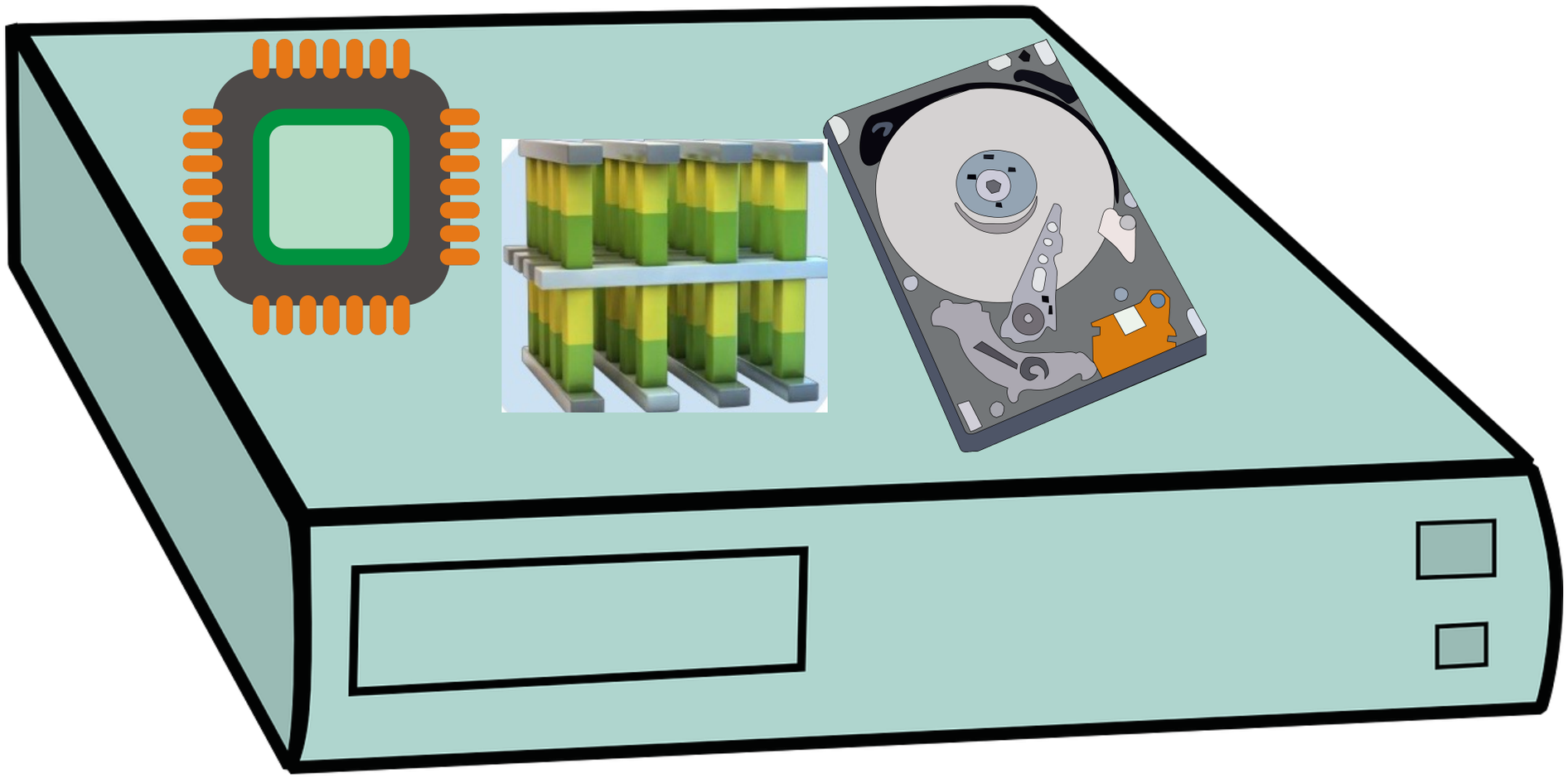
App App App App

OS/Hypervisor



App App App App

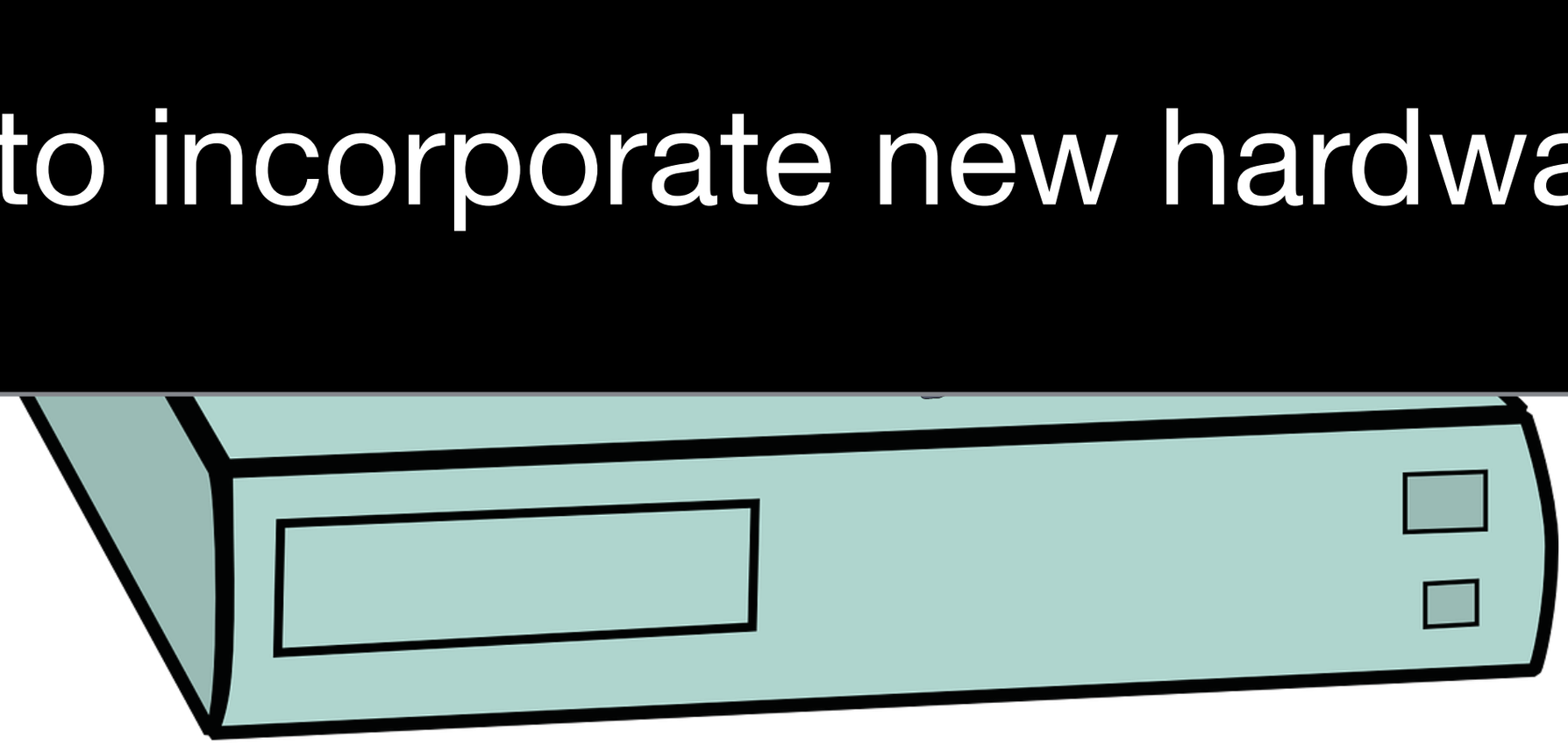
OS/Hypervisor



App App App App

OS/Hypervisor

Difficult to incorporate new hardware



Monolithic Server

- Resource utilization
- Failure
- Flexibility
- Memory capacity wall

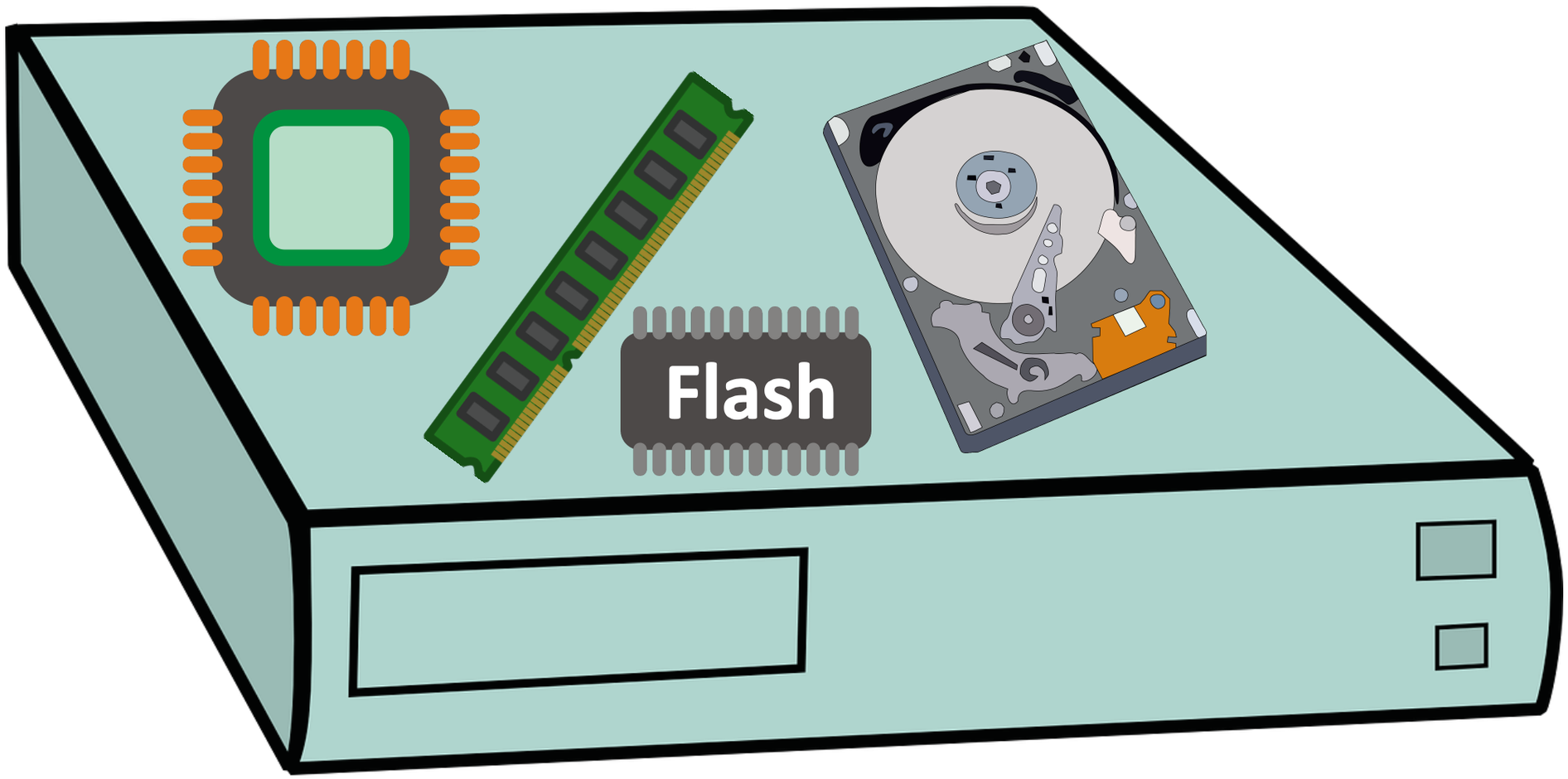
When was the last time

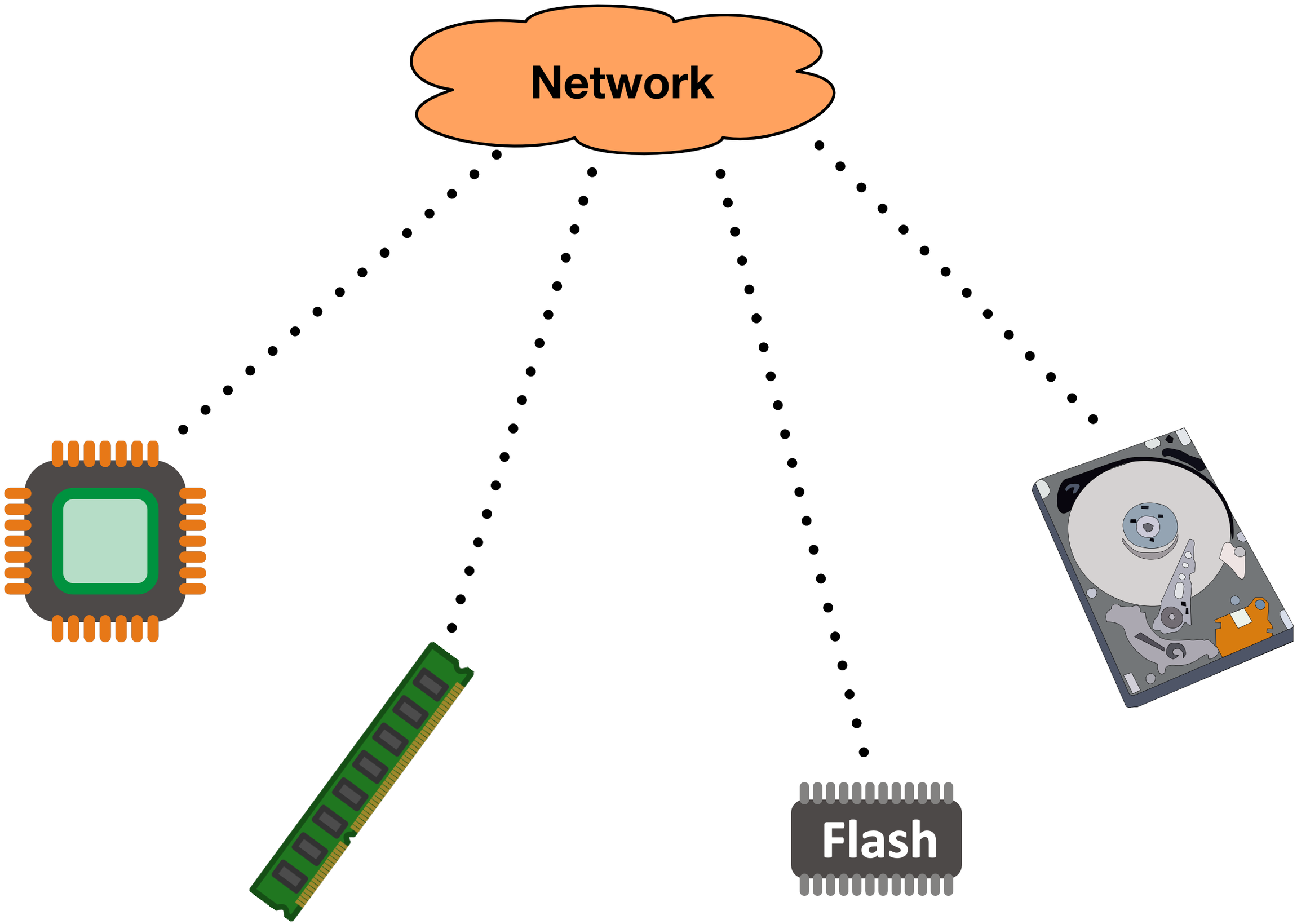
you did something for the first time?


TIME FOR CHANGE

Resource Disaggregation:

Breaking monolithic servers into network-attached, independent hardware components

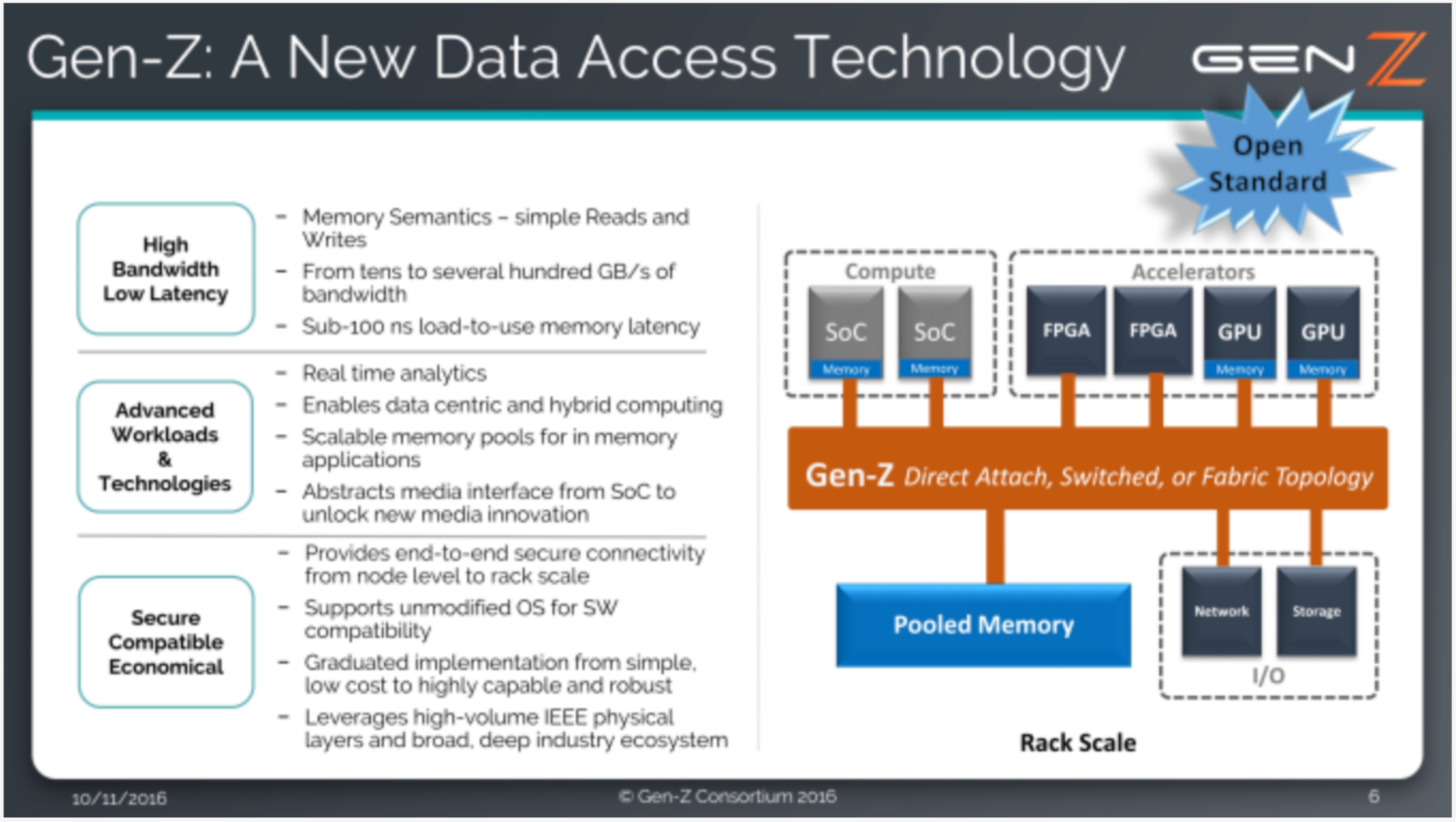

Flash


Network
Flash

Gen-Z Consortium Formed: Developing a New Memory Interconnect

by Ian Cutress on October 12, 2016 9:30 AM EST

Posted in SoC | Samsung | ARM | Huawei | Broadcom | IBM | Mellanox | Xilinx | Interconnect | Gen-Z | HPE



Gen-Z Consortium Formed: Developing a New Memory Interconnect

by Ian Cutress **Oct 2016** 9:30 AM EST

Posted in SoC | Samsung | ARM | Huawei | Broadcom | IBM | Mellanox | Xilinx | Interconnect | Gen-Z | HPE

Gen-Z: A New Data Access Technology

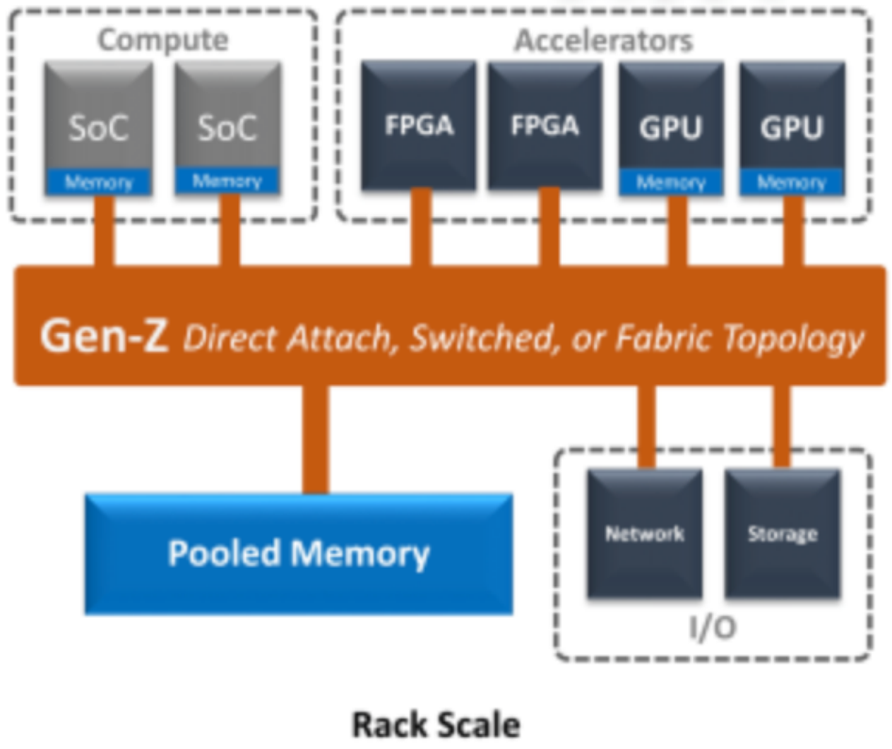
GEN Z

Open Standard

High Bandwidth Low Latency
- Memory Semantics – simple Reads and Writes
- From tens to several hundred GB/s of bandwidth
- Sub-100 ns load-to-use memory latency

Advanced Workloads & Technologies
- Real time analytics
- Enables data centric and hybrid computing
- Scalable memory pools for in memory applications
- Abstracts media interface from SoC to unlock new media innovation

Secure Compatible Economical
- Provides end-to-end secure connectivity from node level to rack scale
- Supports unmodified OS for SW compatibility
- Graduated implementation from simple, low cost to highly capable and robust
- Leverages high-volume IEEE physical layers and broad, deep industry ecosystem



Rack Scale

10/11/2016 © Gen-Z Consortium 2016 6

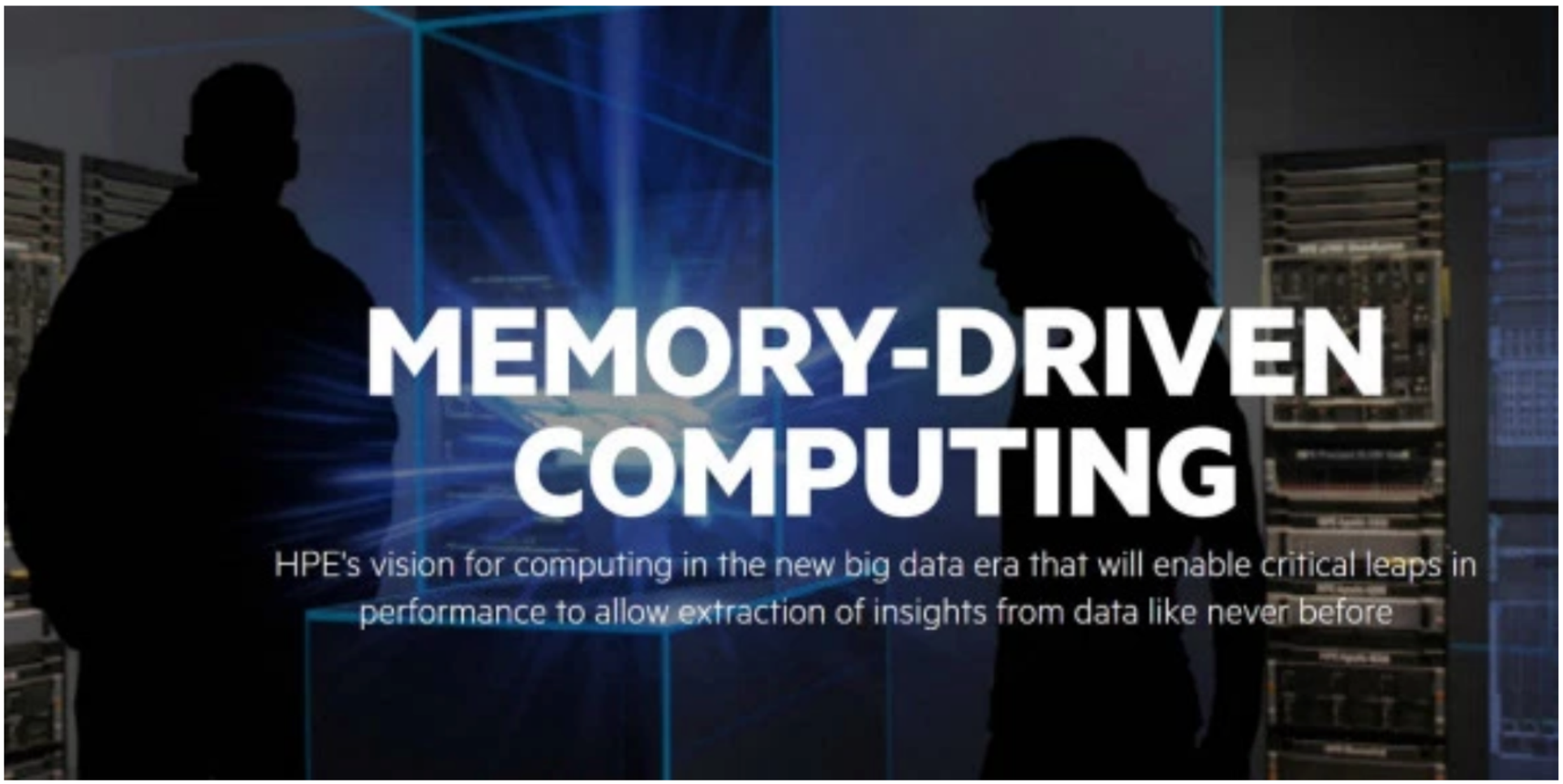
Gen-Z Consortium Formed: Developing a New Memory Interconnect

by Ian Cut

Posted in SoC

HP Enterprise unveils The Machine, a single-memory computer capable of addressing 160 terabytes

DEAN TAKAHASHI @DEANTAK MAY 16, 2017 6:01 AM



Above: HPE's new Memory-Driven Computer puts memory, not the processor, at the center.
Image Credit: HPE

Gen-Z Consortium Formed: Developing a New Memory Interconnect

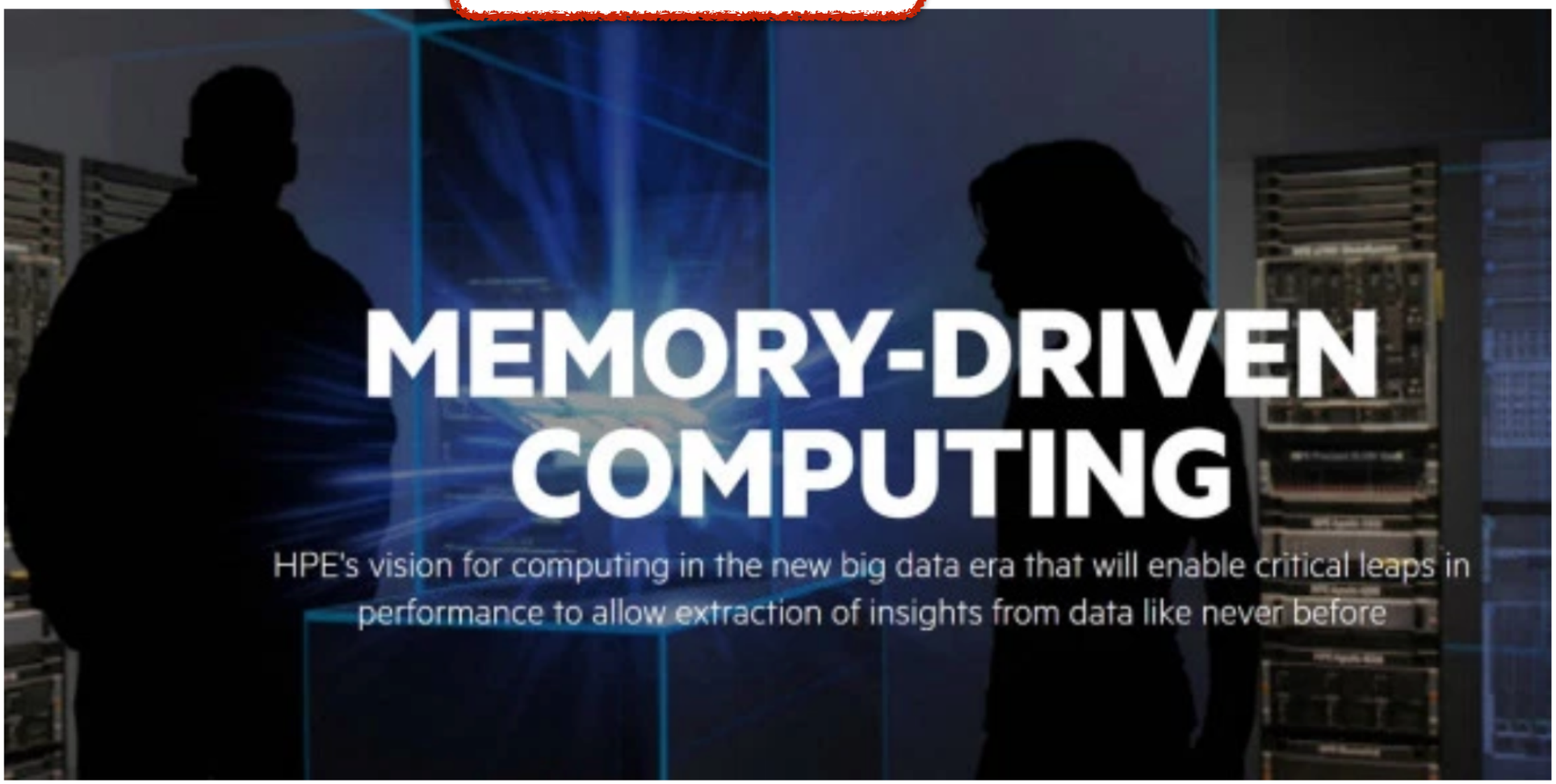
by Ian Cut

Posted in SoC

HP Enterprise unveils The Machine, a single-memory computer capable of addressing 160 terabytes

DEAN TAKAHASHI @DEANTAK

May 2016



Above: HPE's new Memory-Driven Computer puts memory, not the processor, at the center.
Image Credit: HPE

Gen-Z Consortium Formed: Developing a New Memory Interconnect

by Ian Cut

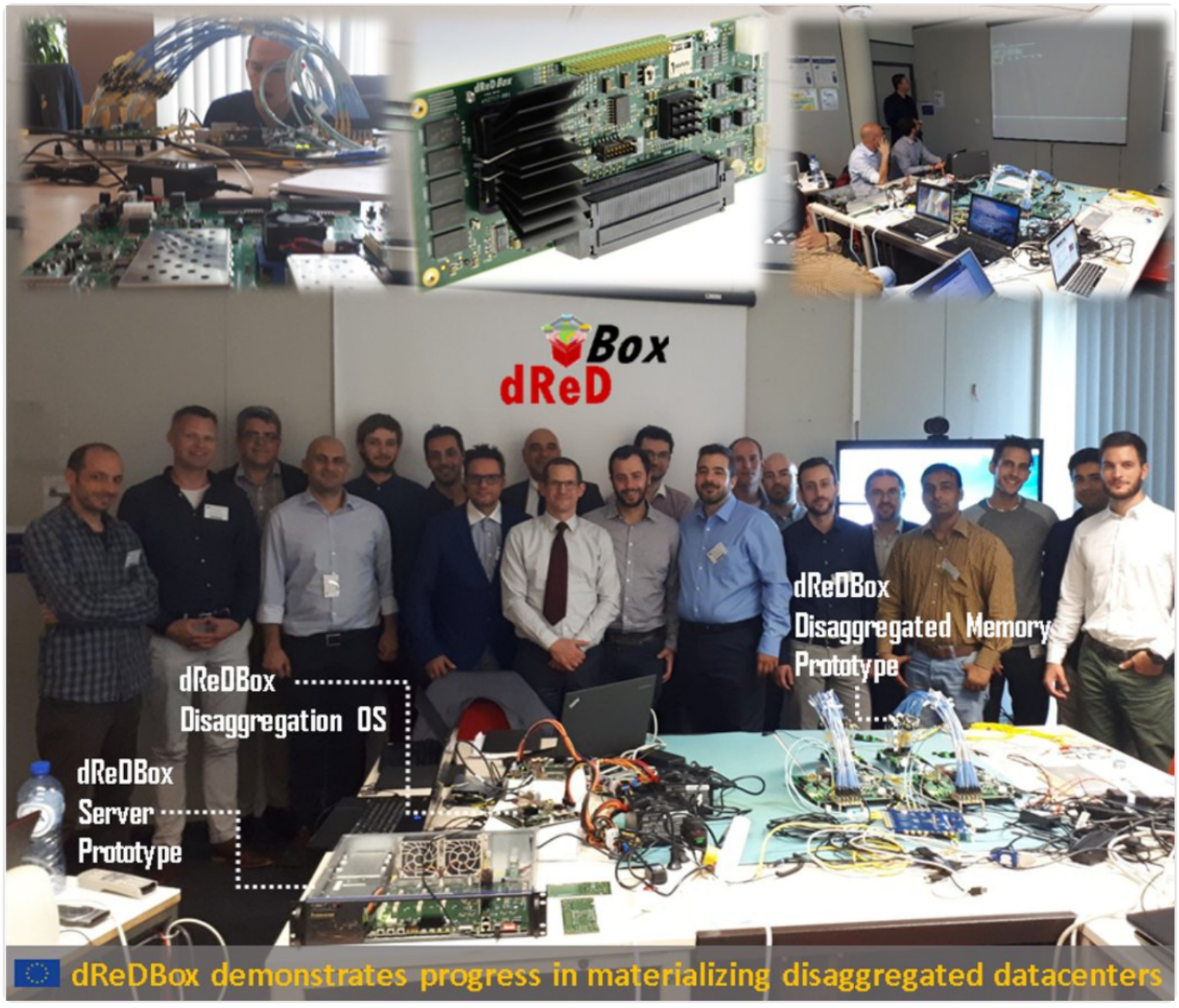
Posted in SoC

HP Enterpris a single-mer of addressin

DEAN TAKAHASHI @DEANTAK

Above: HPE's new Memory-Driven
Image Credit: HPE

dRedBox.eu demonstrates its progress in materializing its vision towards fully disaggregated datacenters and cloud.



6:56 AM - 2 Oct 2017

Gen-Z Consortium Formed: Developing a New Memory Interconnect

by Ian Cut

Posted in SoC

HP Enterpris a single-mer of addressin

DEAN TAKAHASHI @DEANTAK

Above: HPE's new Memory-Driven
Image Credit: HPE

dRedBox.eu demonstrates its progress in materializing its vision towards fully disaggregated datacenters and cloud.

6:56 AM - 2 Oct 2017

Gen-Z Consortium Formed: Developing a New Memory Interconnect

by Ian Cut

Posted in SoC

HP Enterpris a single-mer of addressin

DEAN TAKAHASHI @DEANTAK

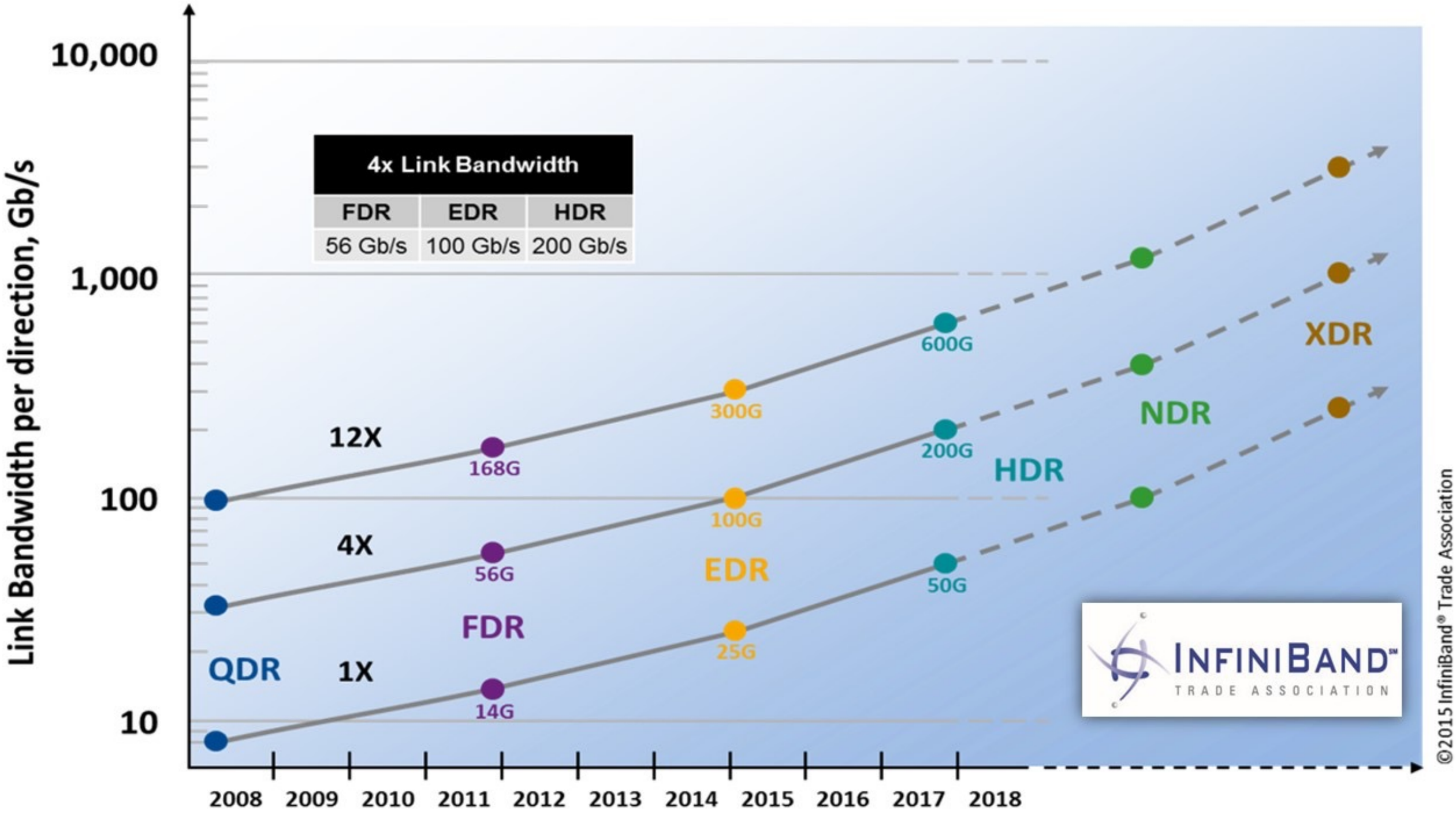
Above: HPE's new Memory-Driven
Image Credit: HPE

dRedBox.eu demonstrates its progress in materializing its vision towards fully disaggregated datacenters and cloud.

Oct 2nd 2017

Why Now?

- Faster network


Link Bandwidth per direction, Gb/s
10,000
1,000
100
10
4x Link Bandwidth
FDR
EDR
HDR
56 Gb/s
100 Gb/s
200 Gb/s
12X
4X
1X
QDR
FDR
EDR
HDR
NDR
XDR
14G
25G
56G
100G
168G
300G
200G
50G
600G
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
INFINIBAND™
TRADE ASSOCIATION
©2015 InfiniBand® Trade Association

Data Center ▸ Networks

Mellanox: We're gonna make InfiniBand great again – 200Gbps great

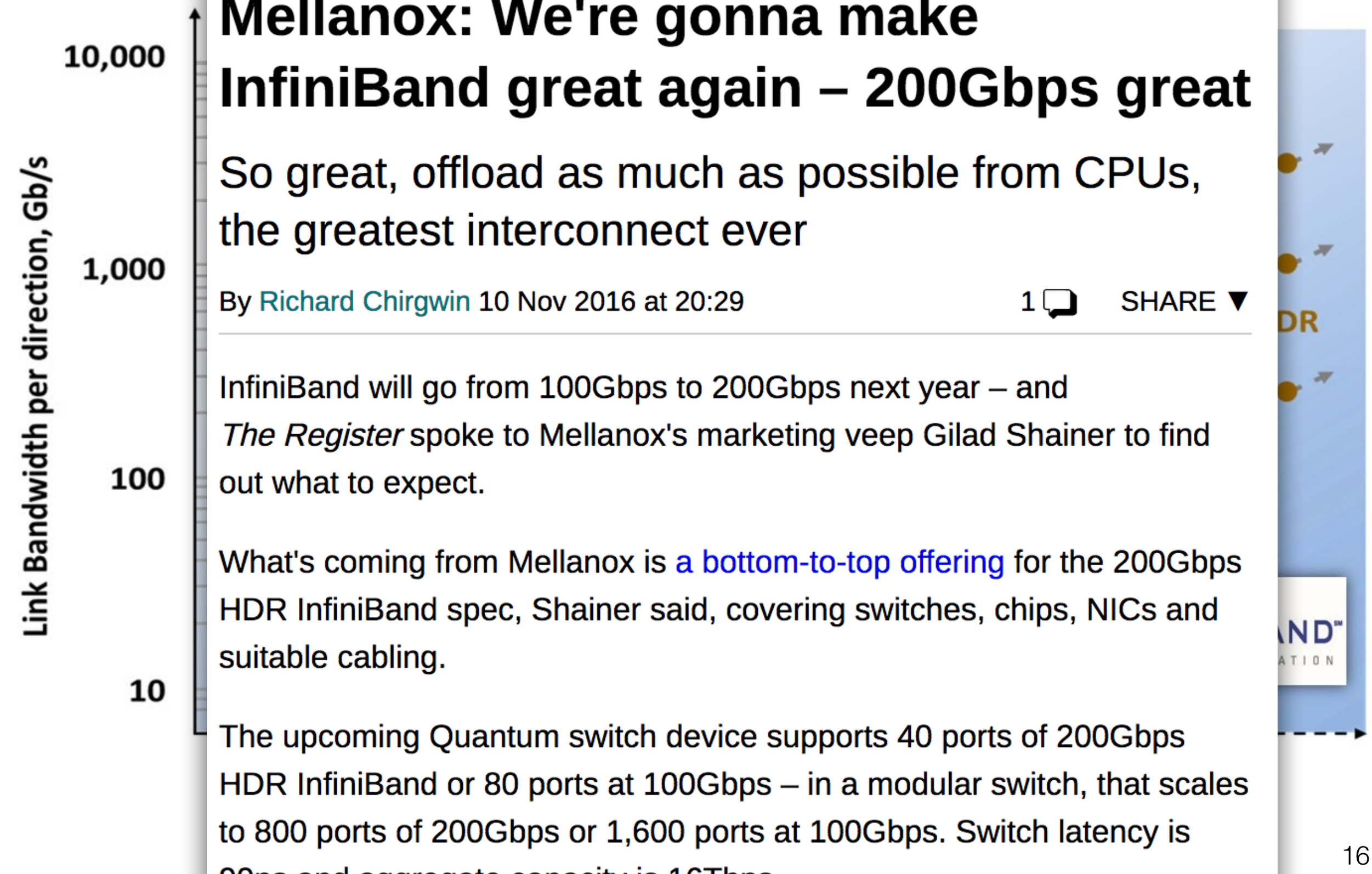
So great, offload as much as possible from CPUs, the greatest interconnect ever

By Richard Chirgwin 10 Nov 2016 at 20:29 1 SHARE ▼

InfiniBand will go from 100Gbps to 200Gbps next year – and *The Register* spoke to Mellanox's marketing veep Gilad Shainer to find out what to expect.

What's coming from Mellanox is a bottom-to-top offering for the 200Gbps HDR InfiniBand spec, Shainer said, covering switches, chips, NICs and suitable cabling.

The upcoming Quantum switch device supports 40 ports of 200Gbps HDR InfiniBand or 80 ports at 100Gbps – in a modular switch, that scales to 800 ports of 200Gbps or 1,600 ports at 100Gbps. Switch latency is 90ns and aggregate capacity is 16Tbps.



Data Center ▸ Networks

Mellanox: We're gonna make InfiniBand great again – 200Gbps great

So great, offload as much as possible from CPUs, the greatest interconnect ever

By Richard Chirgwin 10 Nov 2016 at 20:29 1 SHARE ▼

InfiniBand will go from 100Gbps to 200Gbps next year – and *The Register* spoke to Mellanox's marketing veep Gilad Shainer to find out what to expect.

What's coming from Mellanox is a bottom-to-top offering for the 200Gbps HDR InfiniBand spec, Shainer said, covering switches, chips, NICs and suitable cabling.

The upcoming Quantum switch device supports 40 ports of 200Gbps HDR InfiniBand or 80 ports at 100Gbps – in a modular switch, that scales to 800 ports of 200Gbps or 1,600 ports at 100Gbps. Switch latency is 90ns and aggregate capacity is 16Tbps.

Data Center ▸ Networks

Mellanox: We're gonna make InfiniBand great again - 200Gbps great

So great, offload as much as possible from CPUs,

10,000

Gb/s

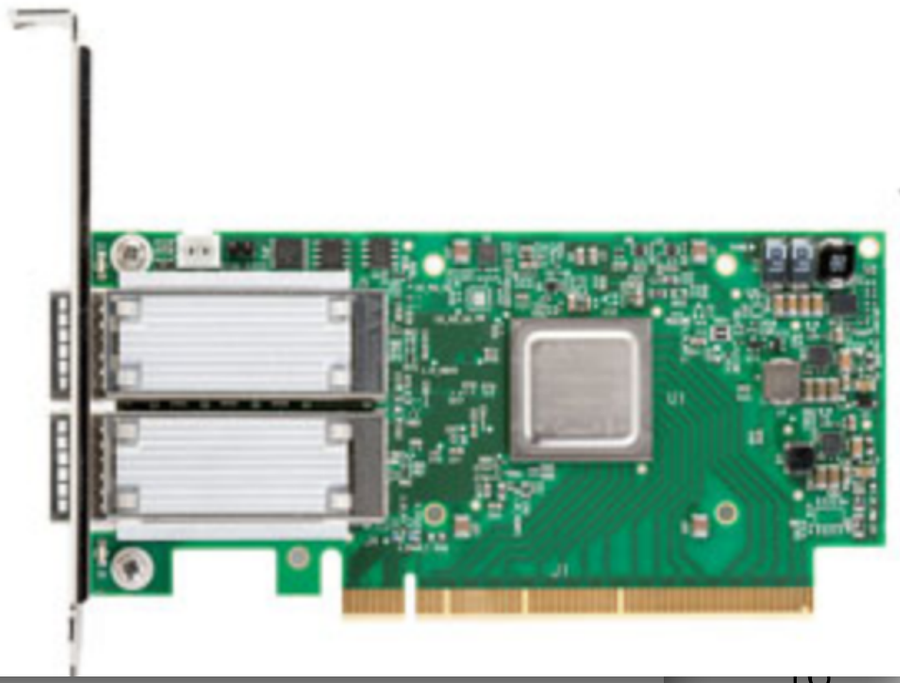
ConnectX®-6 Single/Dual-Port Adapter Supporting 200Gb/s Ethernet

Contact Sales for Availability

Intelligent ConnectX-6 adapter cards, the newest additions to the Mellanox Smart Interconnect suite and supporting Co-Design and In-Network Compute, introduce new acceleration engines for maximizing Cloud, Web 2.0, Big Data, Storage and Machine Learning applications.

ConnectX-6 EN supports two ports of 200Gb/s Ethernet connectivity, sub-600 nanosecond latency, and 200 million messages per second, providing the highest performance and most flexible solution for the most demanding applications and markets.

ConnectX-6 offers Mellanox Accelerated Switching And Packet Processing (ASAP2) Direct technology to offload the vSwitch/vRouter by handling the data plane in the NIC hardware while maintaining the



90ns and aggregate capacity is 16Tbps.

Why Now?

- Faster network
- More powerful hardware controller

Why Now?

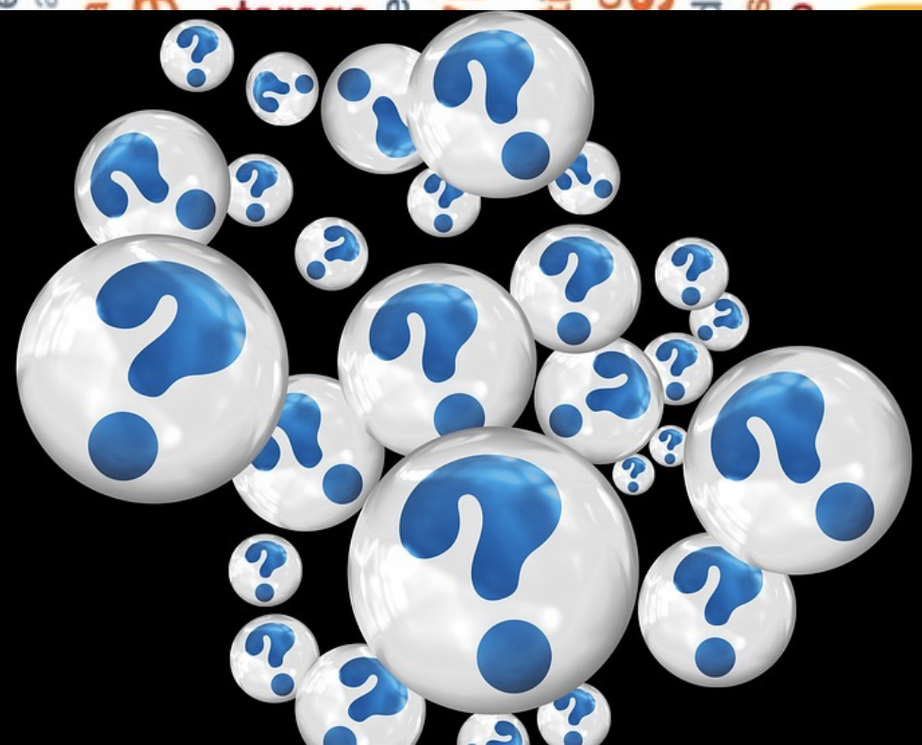
- Faster network
- More powerful hardware controller
- Dynamic application resource requirement
- Quickly changing, heterogeneous hardware

Resource Disaggregation

- Better resource utilization
- Fine-grained failure
- Heterogeneity
- Embracing hardware innovations


Operating
System
interface
hardware
systems
game
machine
request
large
responsible
like
computer
Almost
activities
interact
gooey
command
type
operation
portable
media
offer
generally
Unix
applications
implemented
appliances
OS
programming
using
SUI
data
host
details
however
embedded
computing
use
programs
calls
kind
disk
device
also
consoles
purposes
run
CLI
contention
players
relieves
number
point
Whether
access
program
parameters
computer
applications
commonly
compact
washing
user
desktop
write
models
receive
Unix-like
software
dishwashers
resources
coordination
contained
part
including
interfaces
manage
invoking
storage
handle
service
one
services
may
users
results
domestic
hand-held
outside
oldest
video
line
sharing
acts
typing
makes
management
pass
pronounced
easier
handheld
commands
well
included
robots
APIs
runs
multi-user
GUI


hardware
Operating
systems
game
machine
request
large
responsible
interact
gooey
Almost
activities
like
computer
acts
typing
makes
command
type
operation
portable
media
offer
generally
applications
Unix
appliances
implemented
programming
System
use
programs
embedded
computing
data
host
details
however
kind
calls
using
purposes
CLI
run
device
also
disk
consoles
contention
players
relieves
parameters
number
washing
including
interfaces
manage
invoking
may
user
APIs
runs
multi-user
GUI
sharing
management
pass
handheld
commands
well
included
easier
pronounced
robots



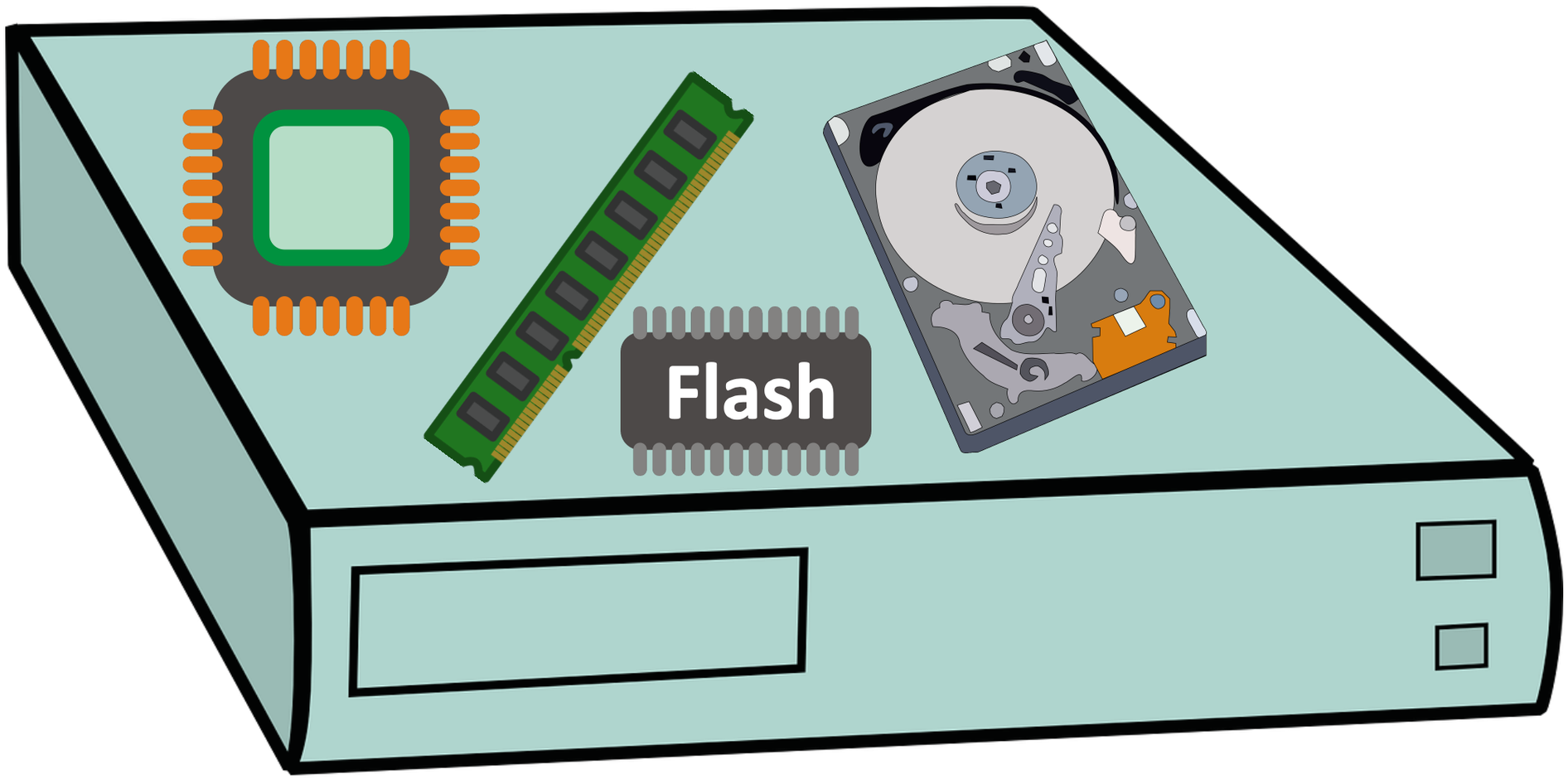
Using Existing Kernels?

- Monolithic/micro kernel: built for single monolithic server
- Multikernel: (vertically) replicated kernel across cores
- Distributed OS [*Sprite*, *V*, *MOSIX*, *Charlotte*]:

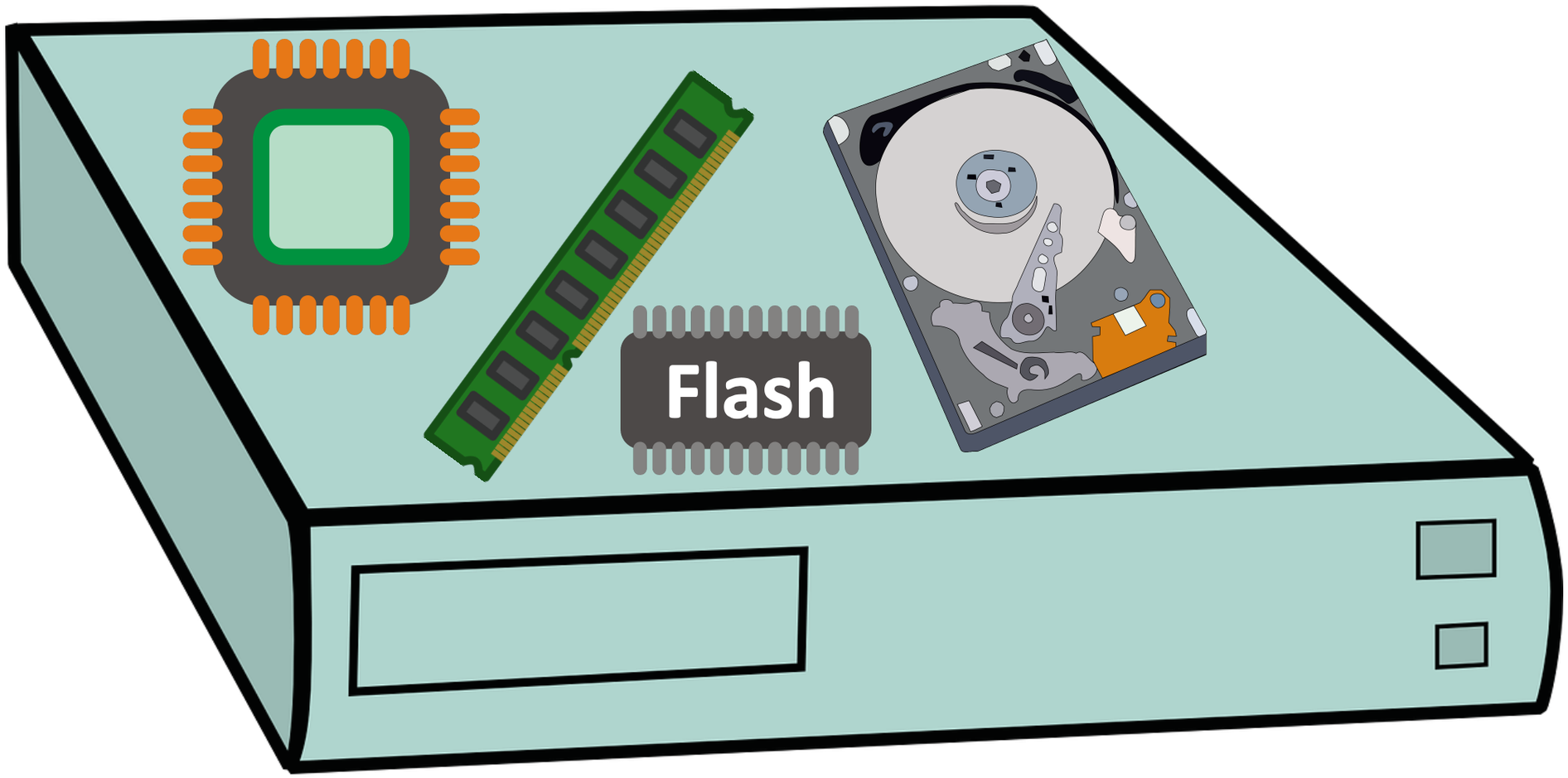
  manages distributed monolithic servers

  Amoeba: manages resource pool, but not in modern days

When hardware is disaggregated, the OS should be also!


Flash

OS



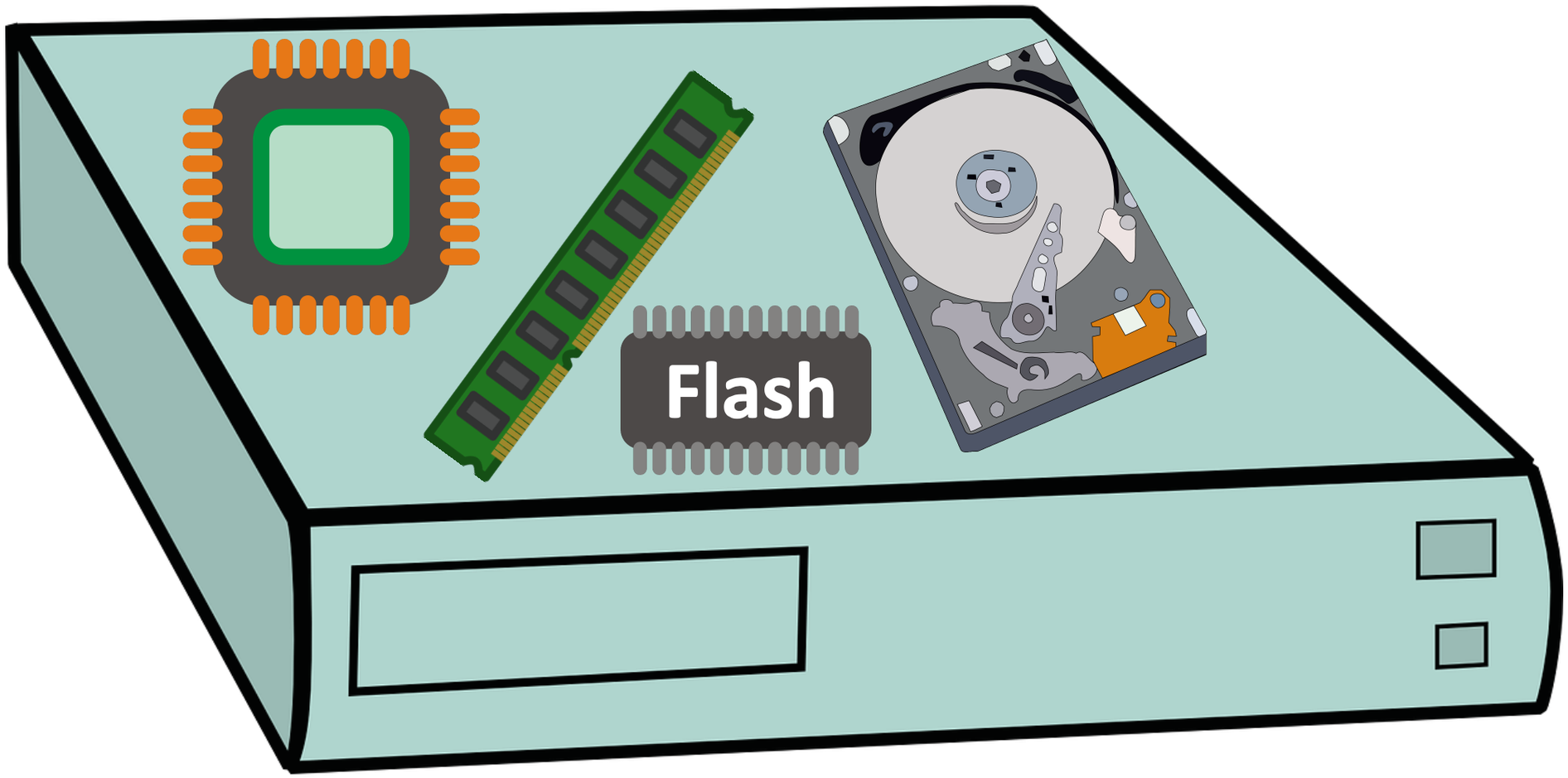
OS

Process Mgmt

Virtual Memory System

File & Storage System

Network



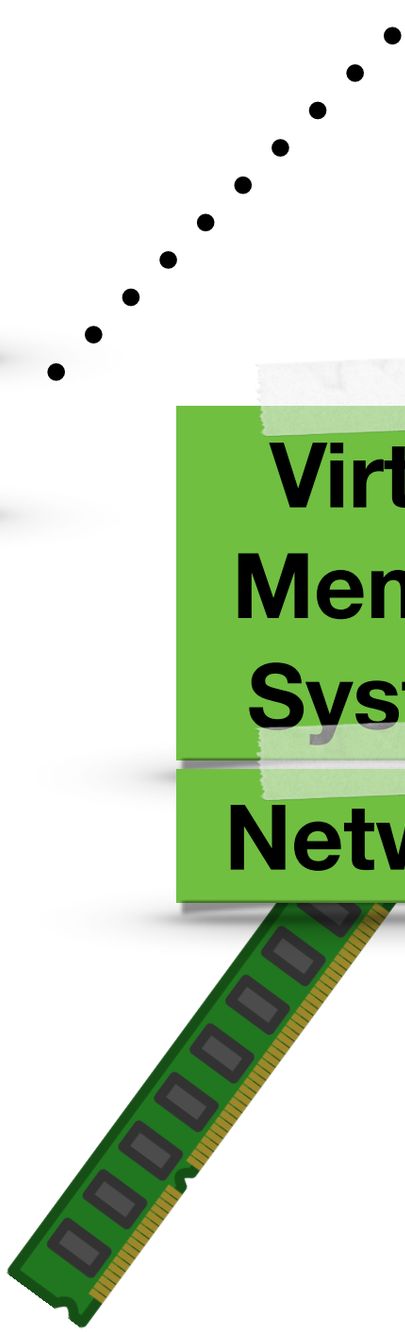
Network

Process Mgmt

Network

Virtual Memory System

Network



File & Storage System

Network

File & Storage System

Network

Flash

Lego

Challenges

- Cleanly separate OS services
  - *Stateless, minimal dependencies*
- Fit hardware constraints
  - Processor: no or limited local DRAM
  - Memory: limited processing power

Network

Process Mgmt

Network

Virtual Memory System

Network

File & Storage System

Network

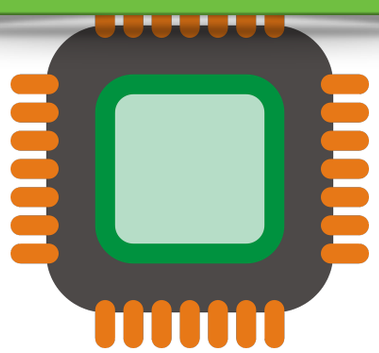
File & Storage System

Network

Flash

Process Mgmt

Network



Network

Virtual Memory System

Network

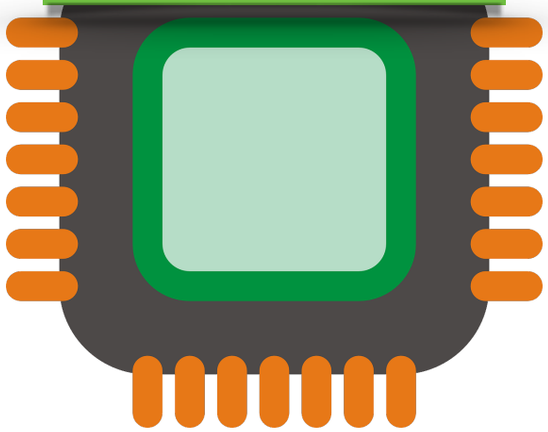
File & Storage System

Network

Flash

Process Mgmt

Network



Network

Virtual Memory System

Network

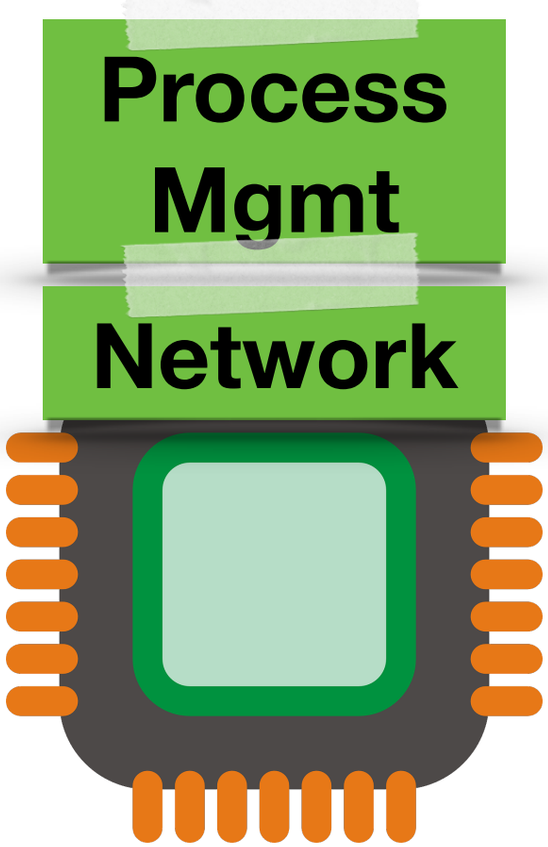
File & Storage System

Network

Flash

Process Mgmt

Network



Network

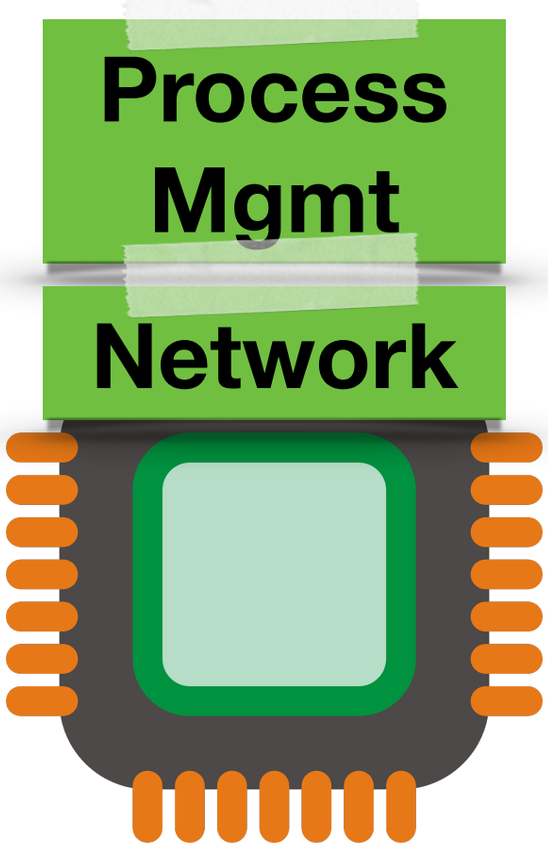
File & Storage System

Network

Flash

Process Mgmt

Network



Network

File & Storage System

Core

L1 $

L2 $

Core

L1 $

L2 $

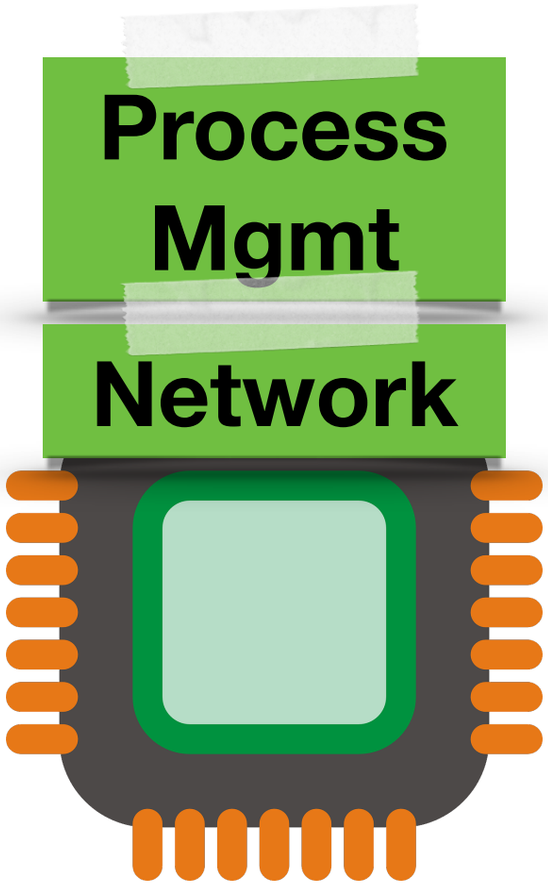
Core

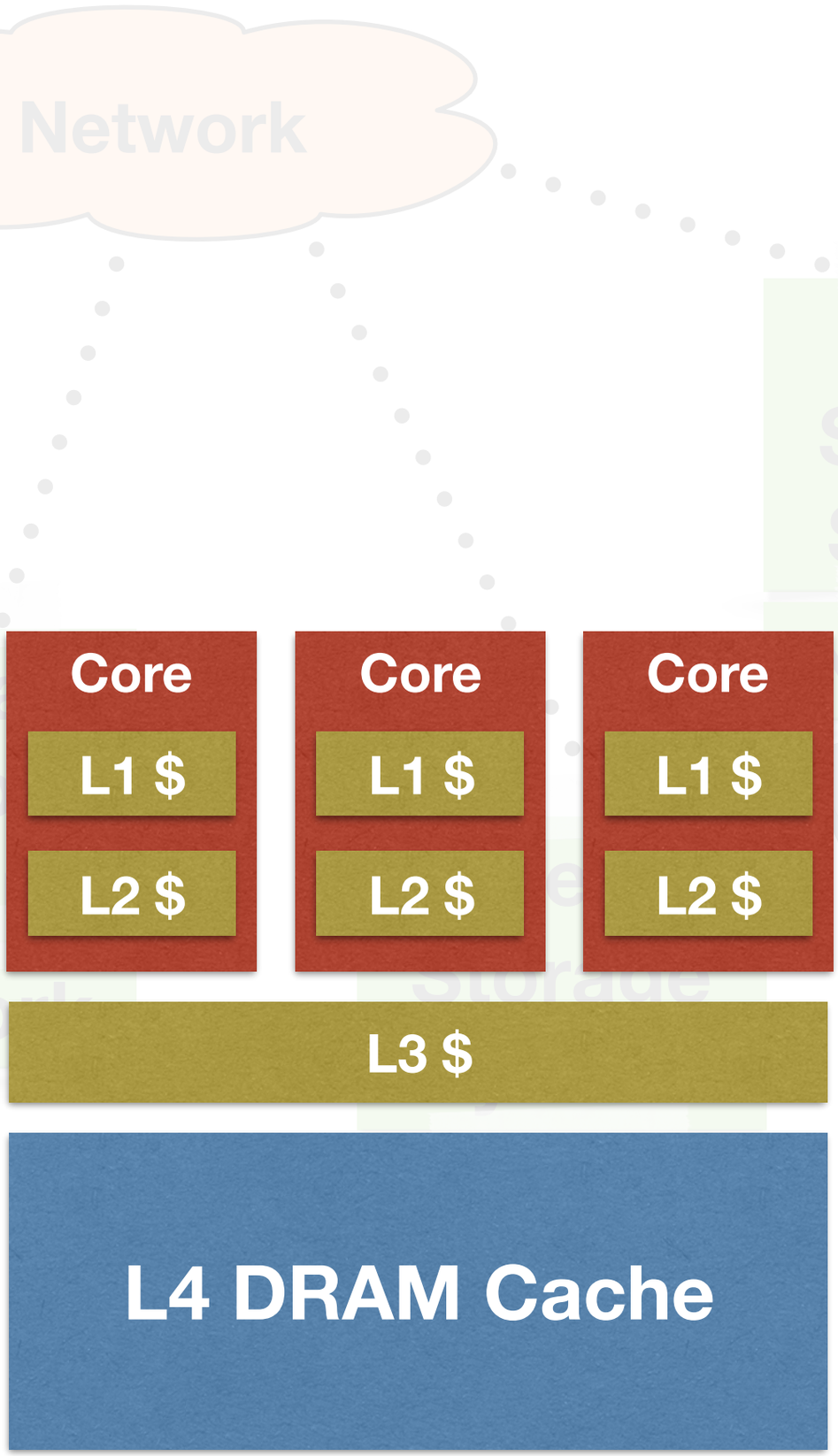
L1 $

L2 $

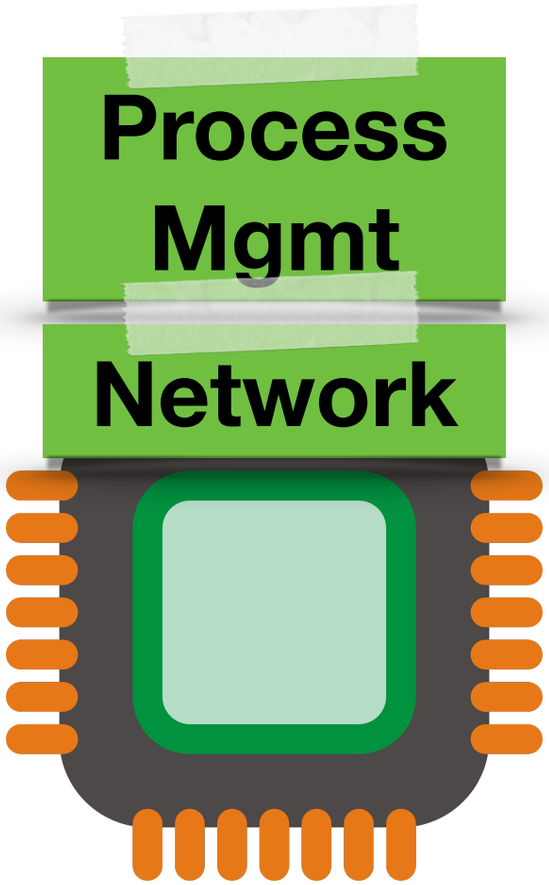
L3 $

L4 DRAM Cache

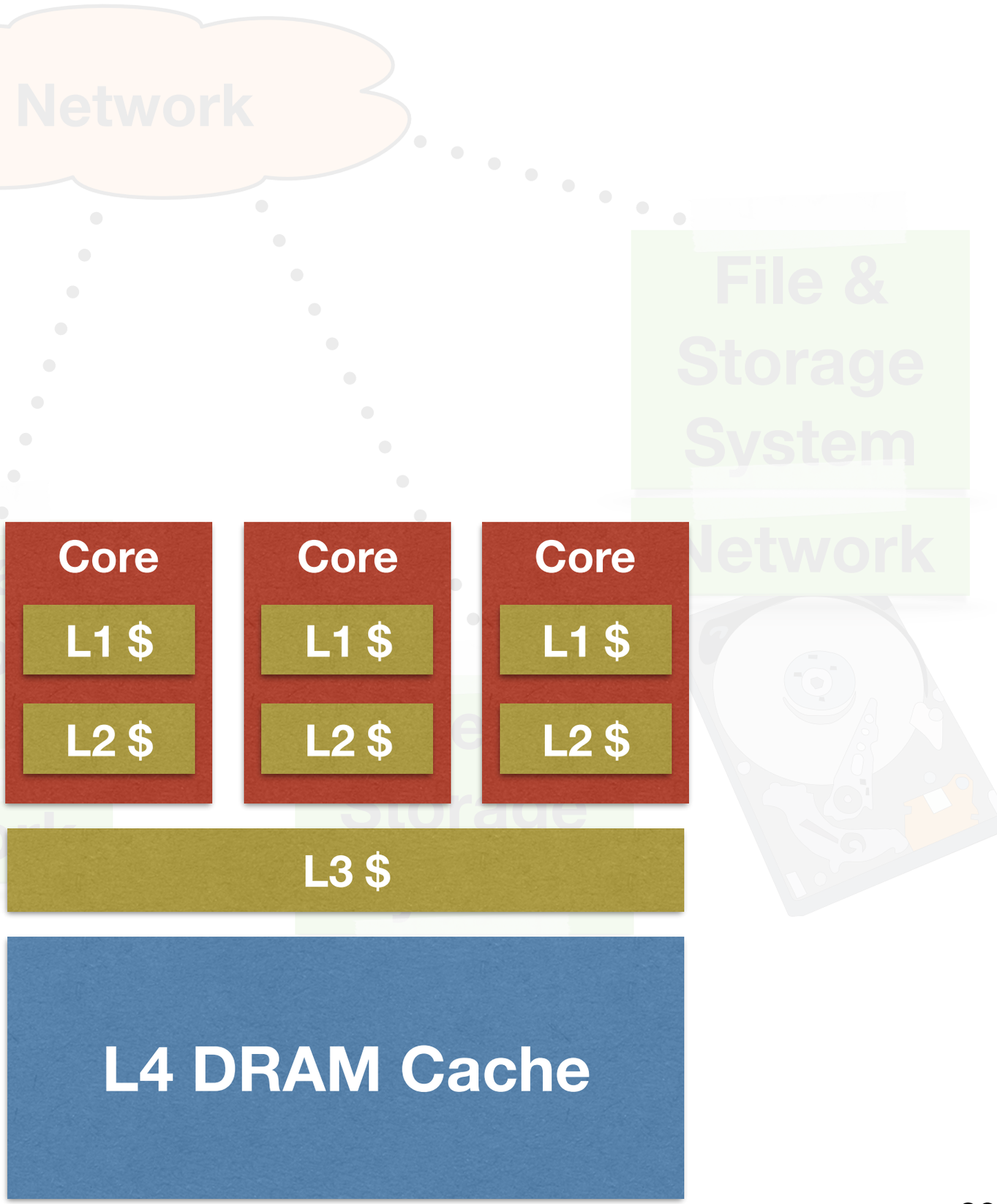


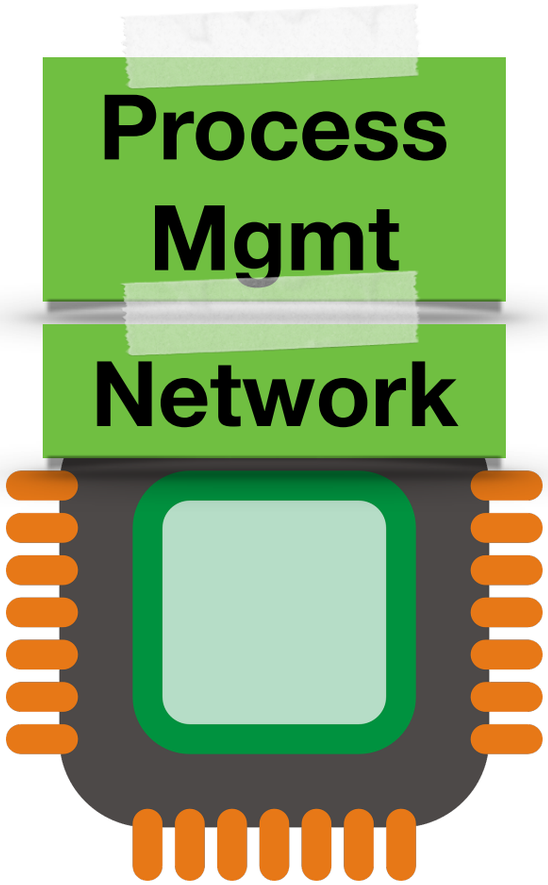


- Software managed

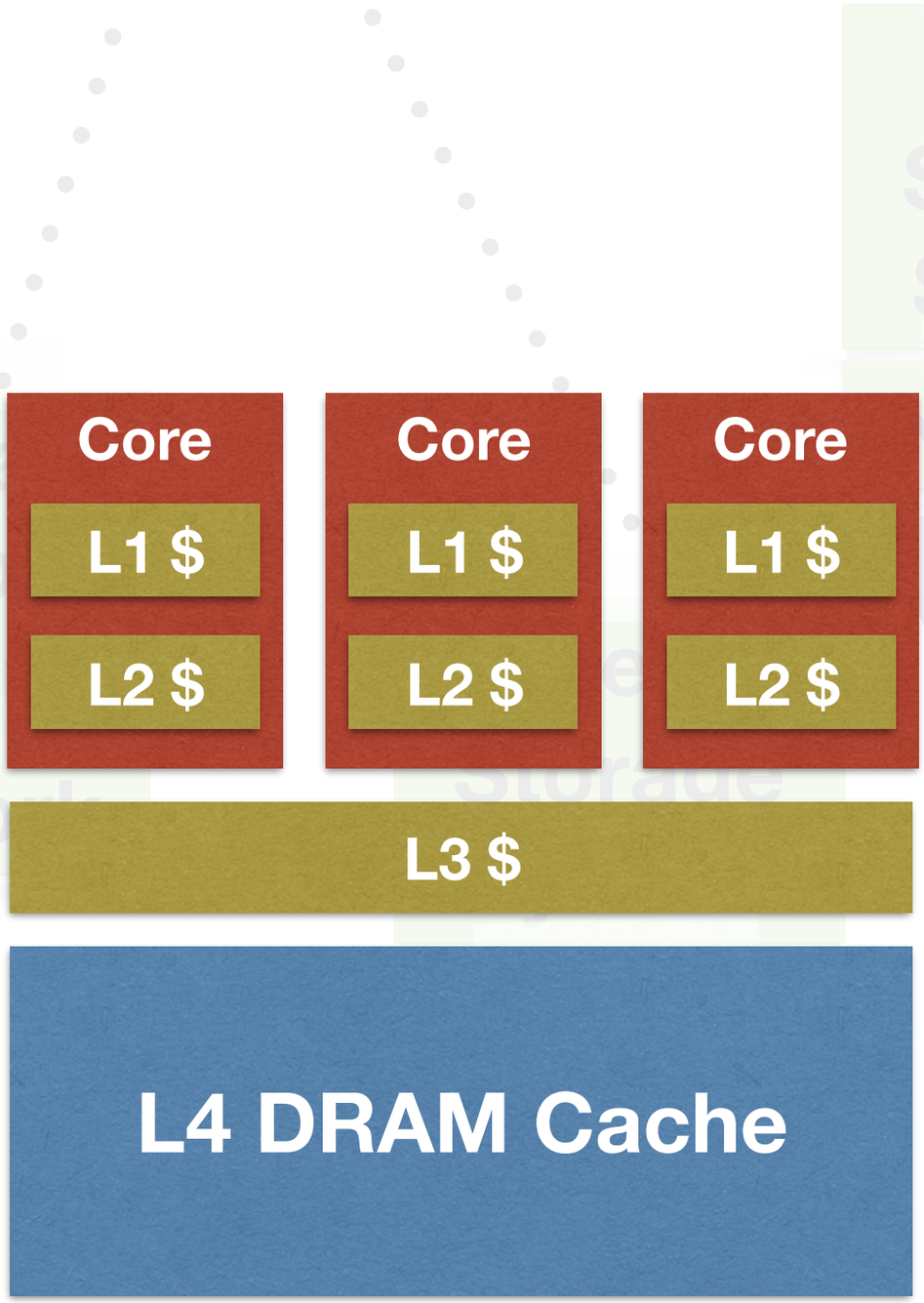


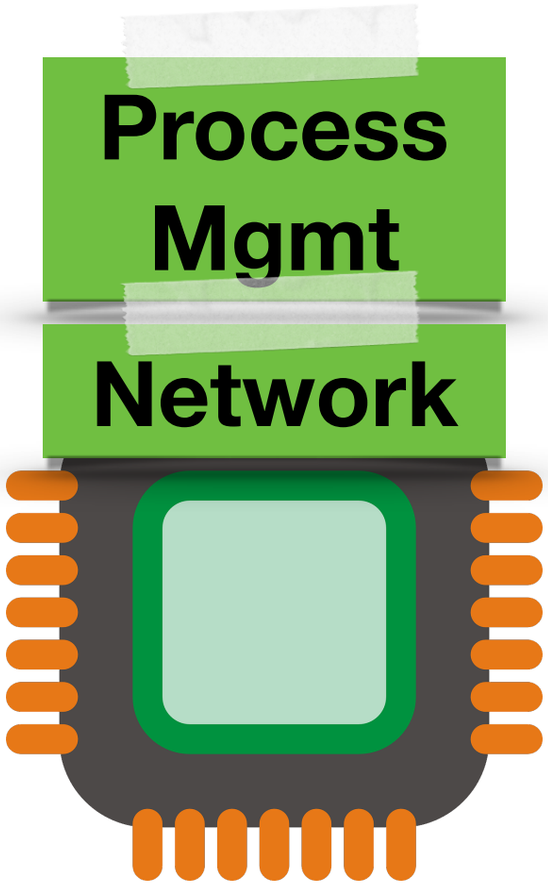
- Coarse $ line
- Software managed

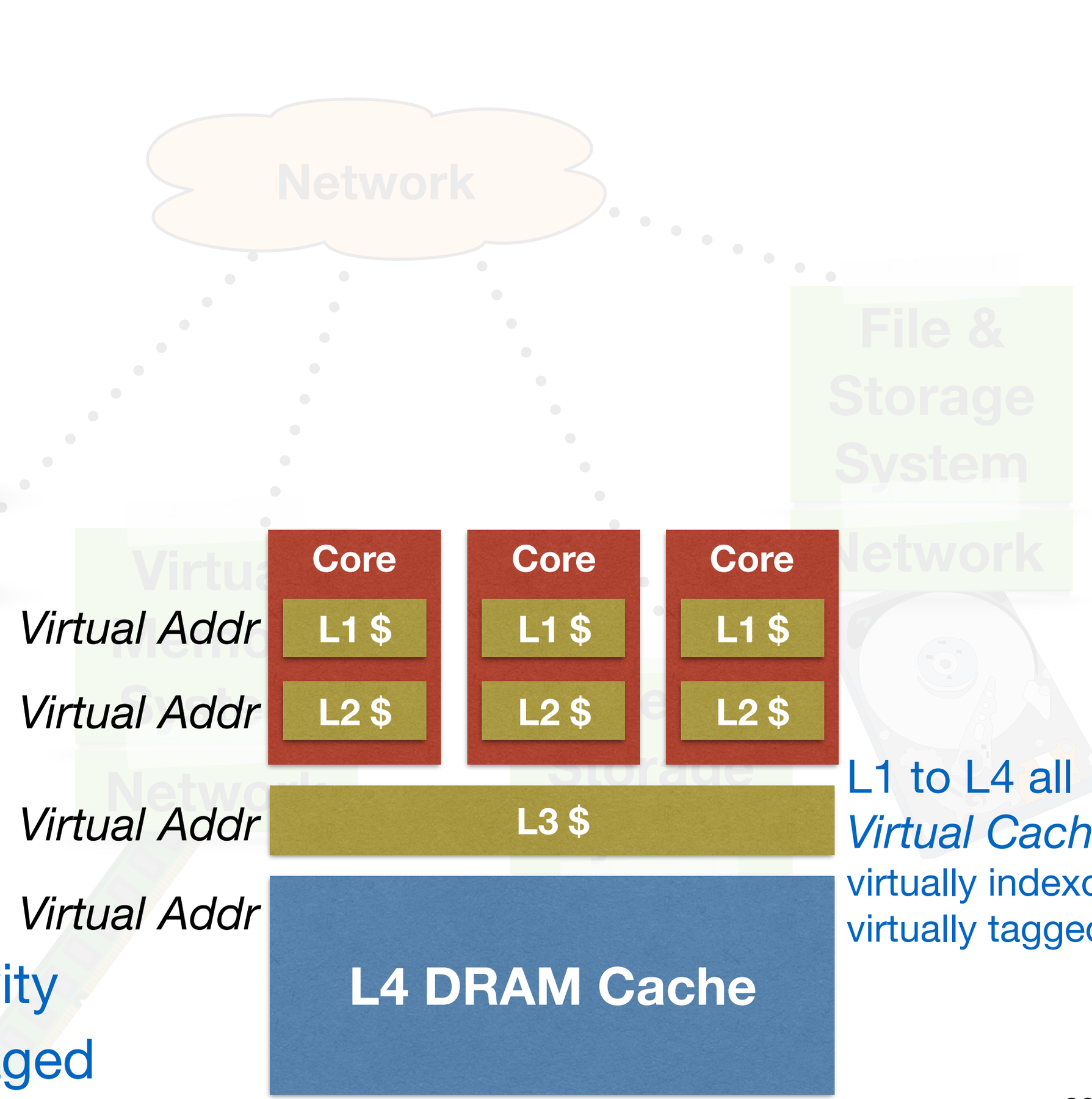




- Coarse $ line
- High associativity
- Software managed







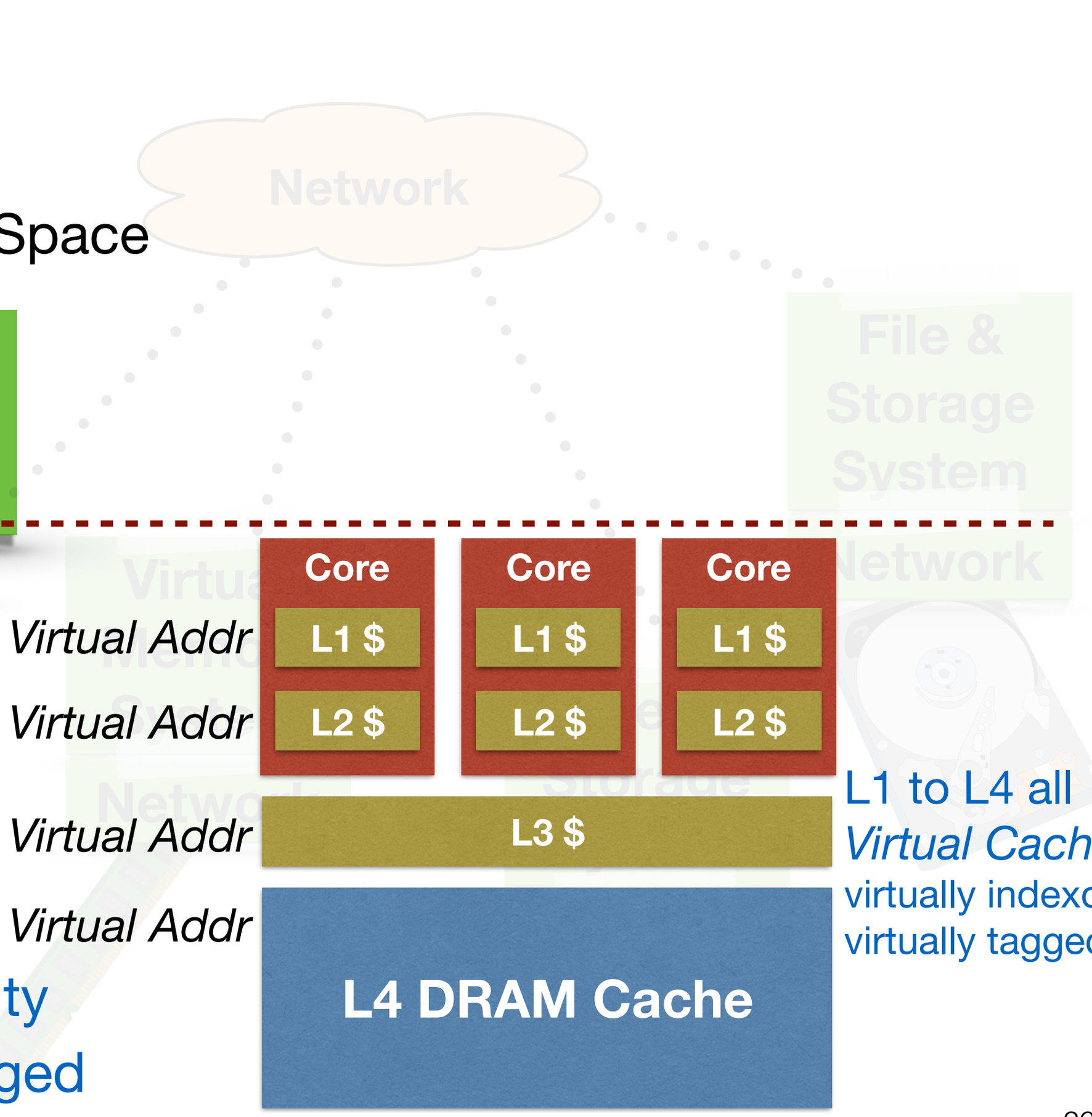
- Coarse $ line
- High associativity
- Software managed

L1 to L4 all *Virtual Cache*:
virtually indexd
virtually tagged

Kernel Space

Process Mgmt

Network



- Coarse $ line
- High associativity
- Software managed

L1 to L4 all *Virtual Cache*:
virtually indexd
virtually tagged

Kernel Space

Process/thread scheduling

Process Mgmt

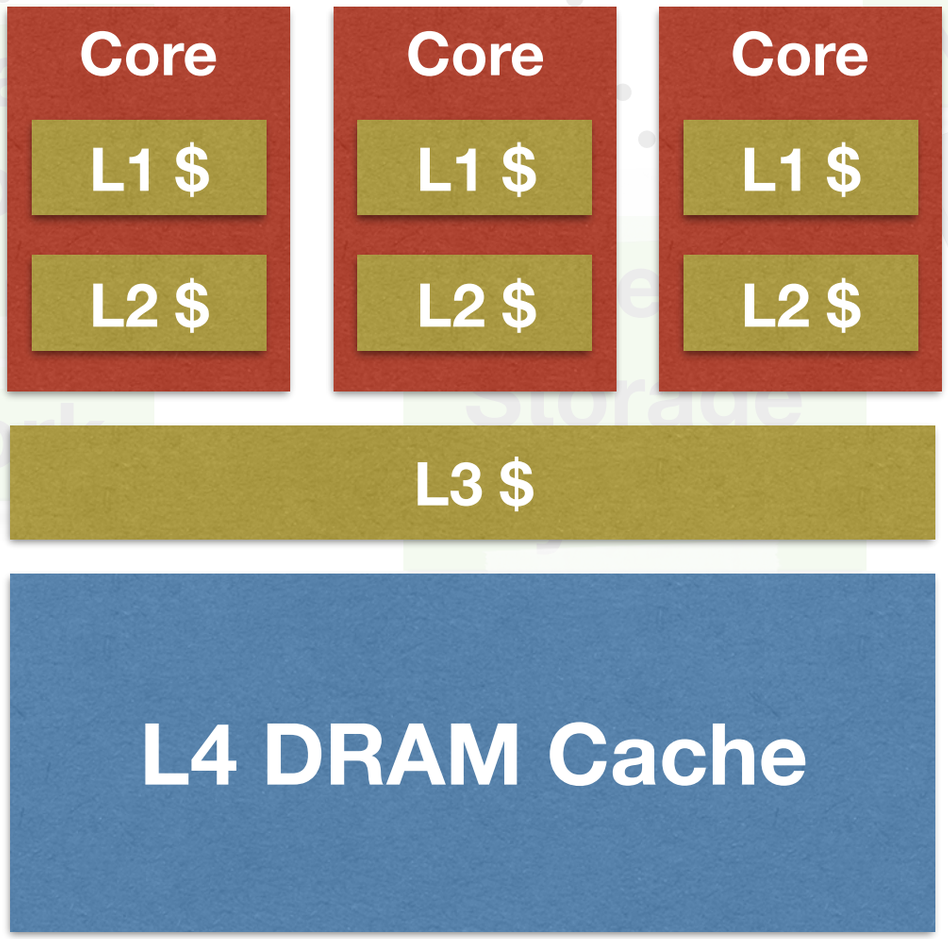
Network

Virtual Addr

Virtual Addr

Virtual Addr

Virtual Addr



L1 to L4 all *Virtual Cache*:
virtually indexd
virtually tagged

- Coarse $ line
- High associativity
- Software managed

Kernel Space

Process/thread scheduling

L4 cache management

Process Mgmt

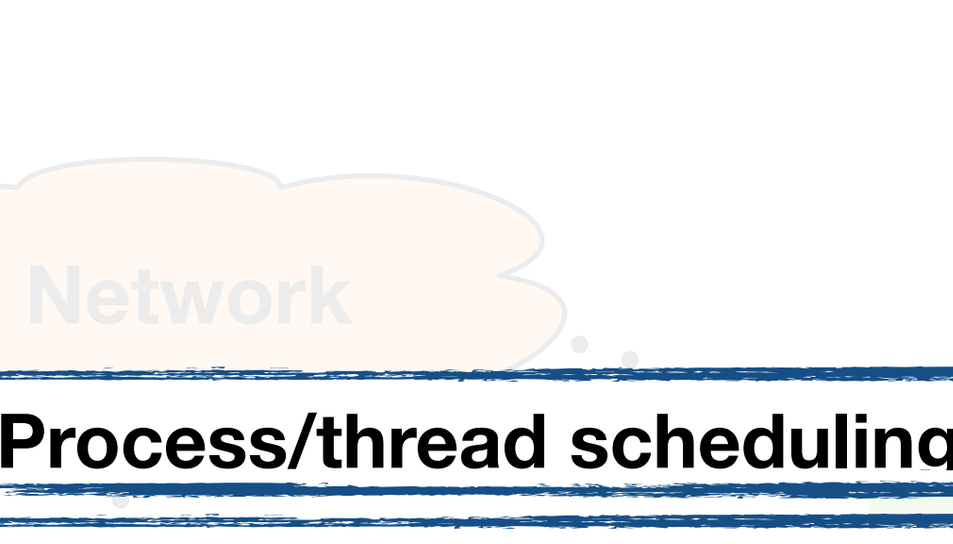
Network

Virtual Addr

Virtual Addr

Virtual Addr

Virtual Addr



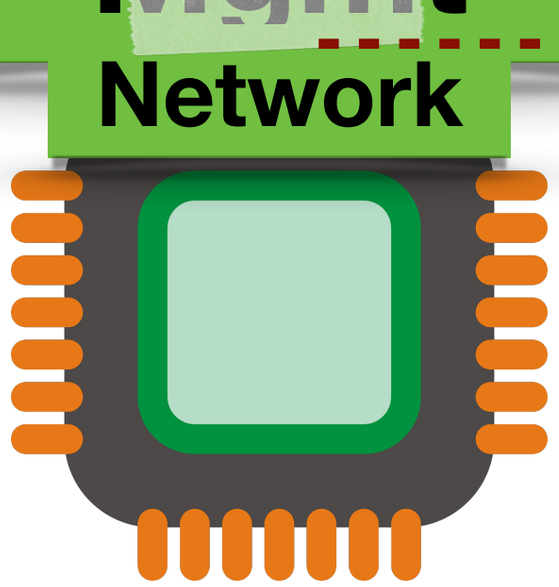
- Coarse $ line
- High associativity
- Software managed

L1 to L4 all *Virtual Cache*:
virtually indexd
virtually tagged

Kernel Space

Process Mgmt

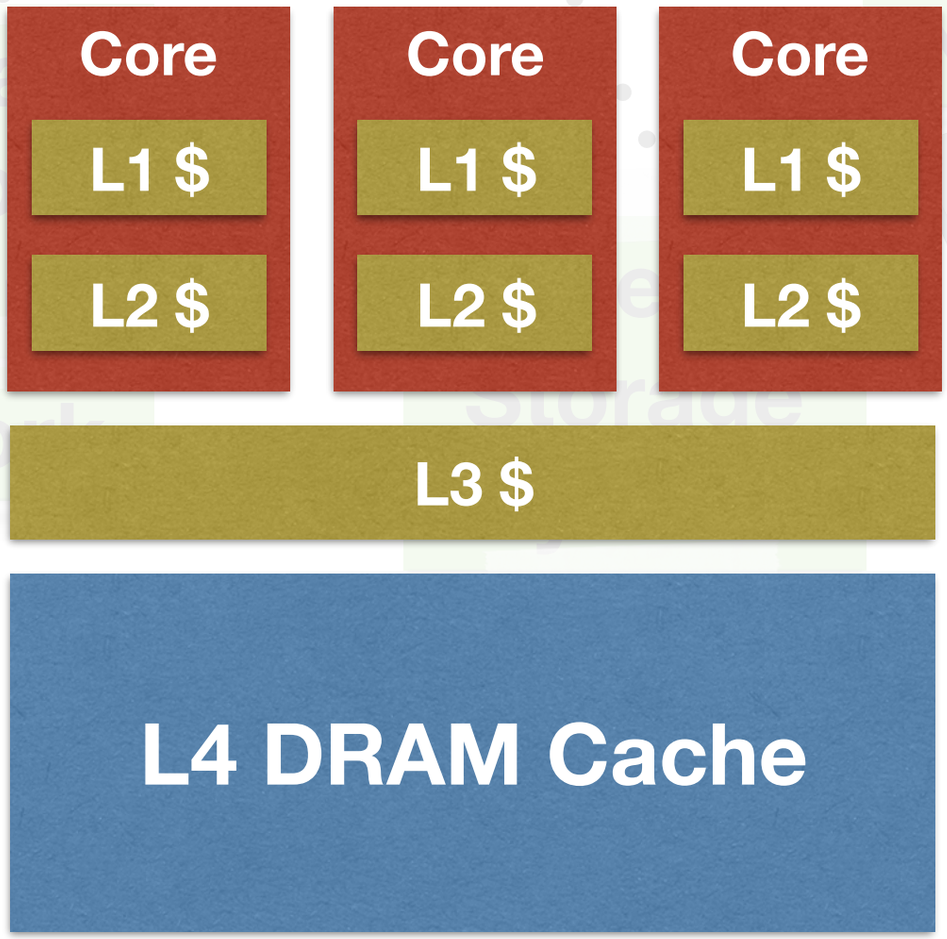
Network



Process/thread scheduling

L4 cache management

Process checkpointing



- Coarse $ line
- High associativity
- Software managed

L1 to L4 all *Virtual Cache*:
virtually indexd
virtually tagged

Kernel Space

Linux interface, state session

Process/thread scheduling

L4 cache management

Process checkpointing

Process Mgmt

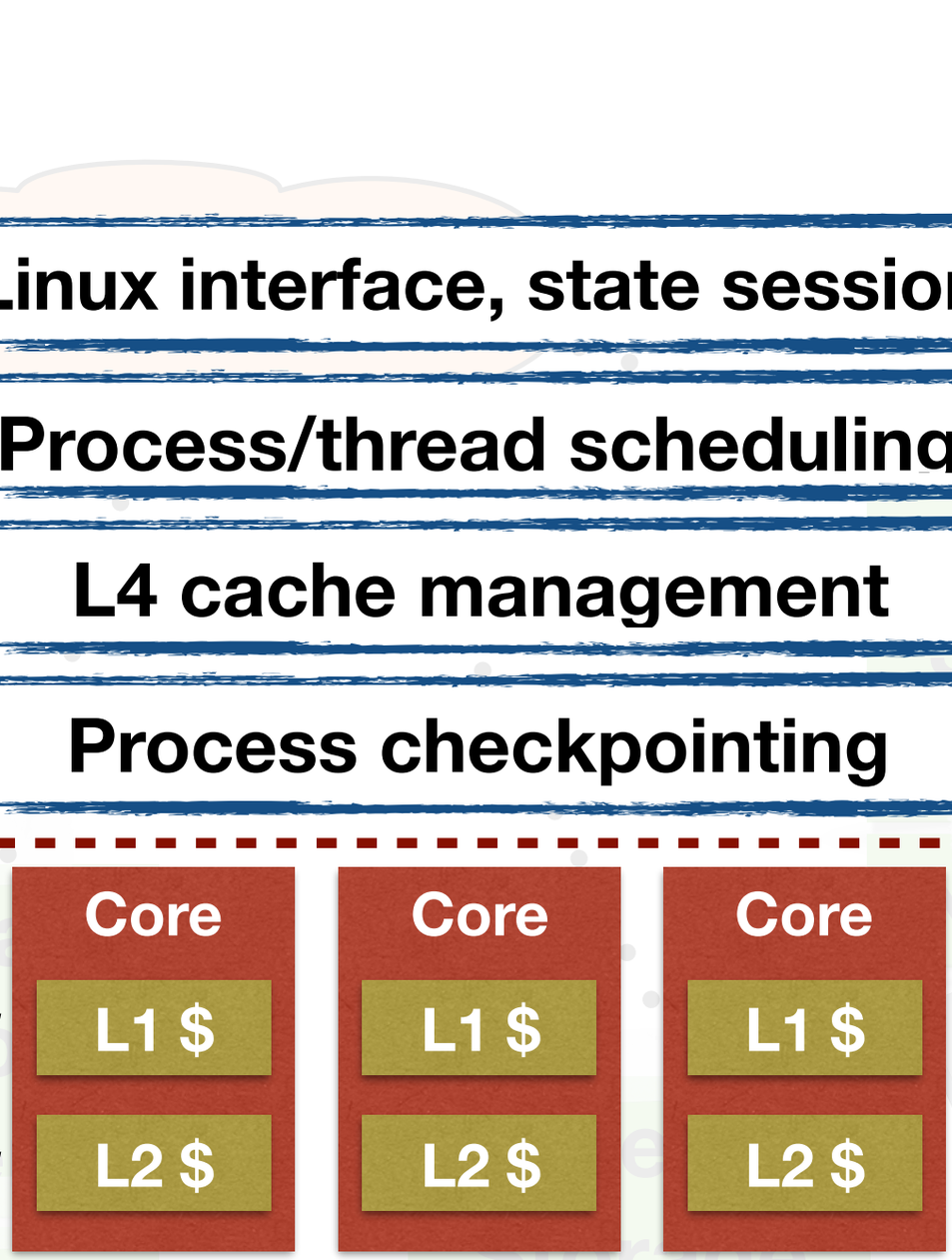
Network

Virtual Addr

Virtual Addr

Virtual Addr

Virtual Addr



L1 to L4 all *Virtual Cache*:
virtually indexd
virtually tagged

- Coarse $ line
- High associativity
- Software managed

User Space

Application processes

Kernel Space

Linux interface, state session

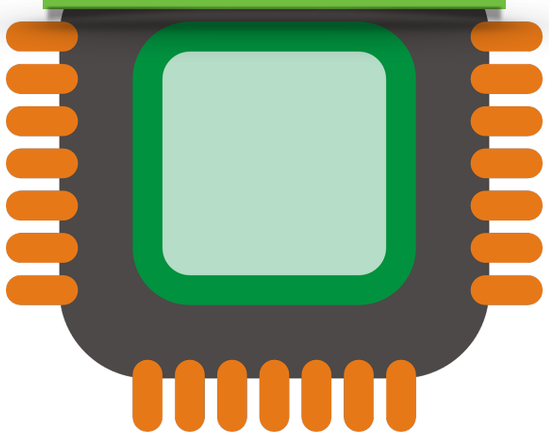
Process/thread scheduling

L4 cache management

Process checkpointing

Process Mgmt

Network

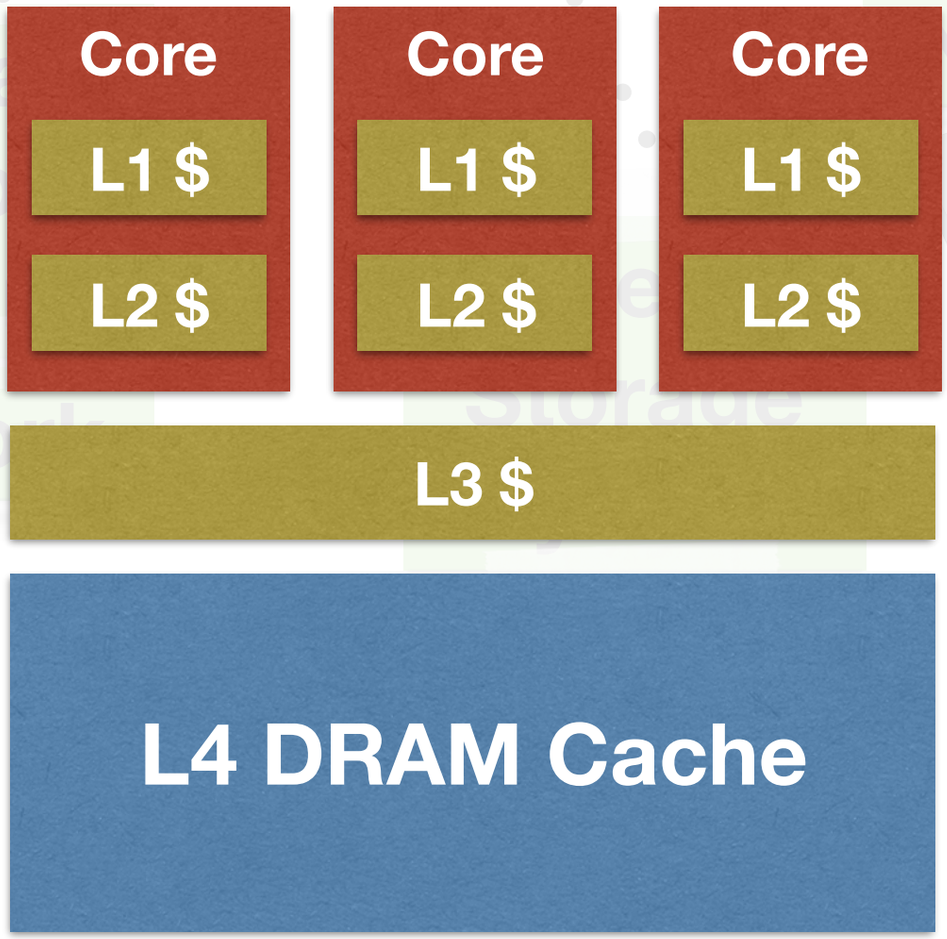


Virtual Addr

Virtual Addr

Virtual Addr

Virtual Addr



L1 to L4 all *Virtual Cache*:
virtually indexd
virtually tagged

- Coarse $ line
- High associativity
- Software managed

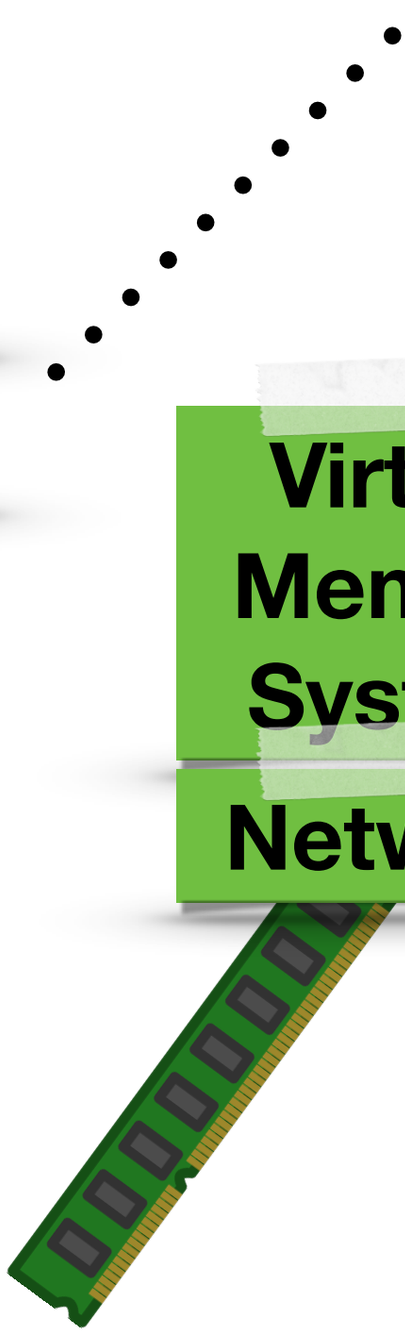
Network

Process Mgmt

Network

Virtual Memory System

Network



File & Storage System

Network

File & Storage System

Network

Flash

Network

Process Mgmt

Network

File & Storage System

Network

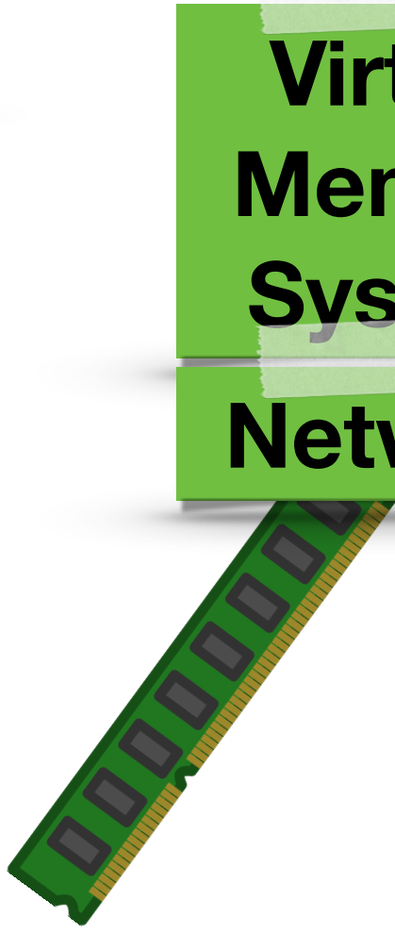
Virtual Memory System

Network

File & Storage System

Network

Flash



Network

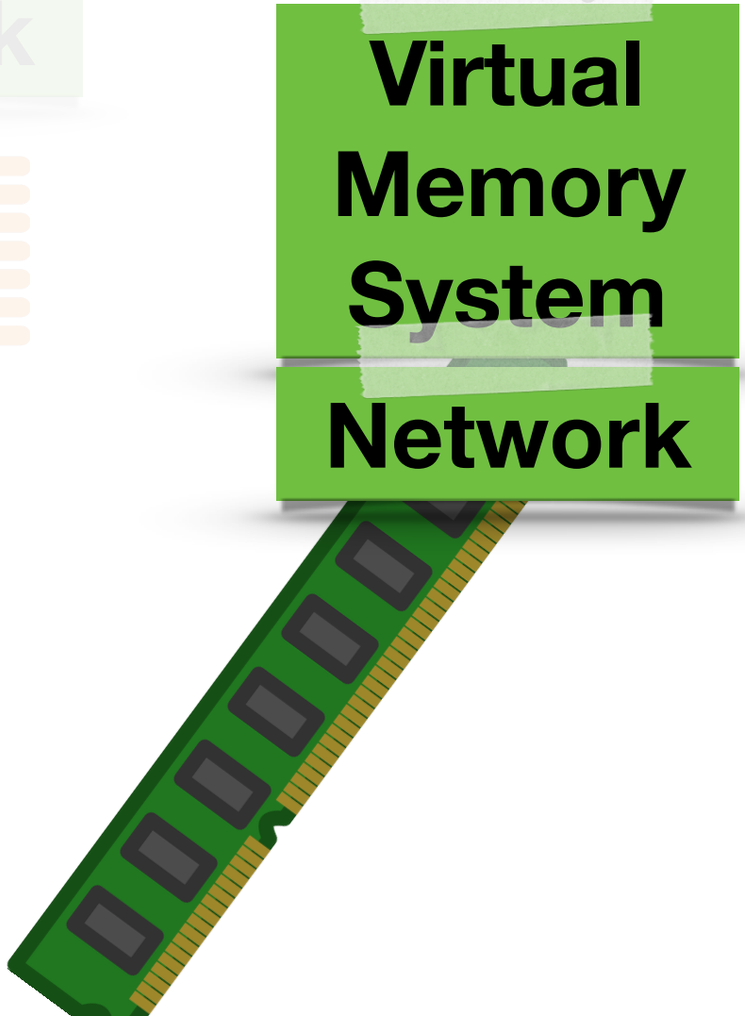
Process Mgmt

File & Storage System

Network

Virtual Memory System

Network



File & Storage System

Network

Flash

Network

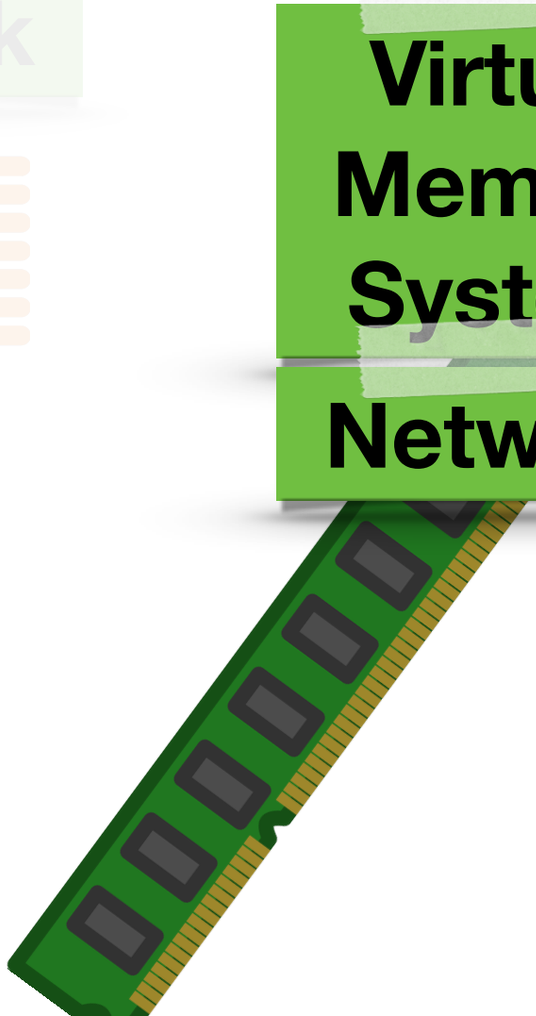
Process Mgmt

File & Storage System

Network

Virtual Memory System

Network



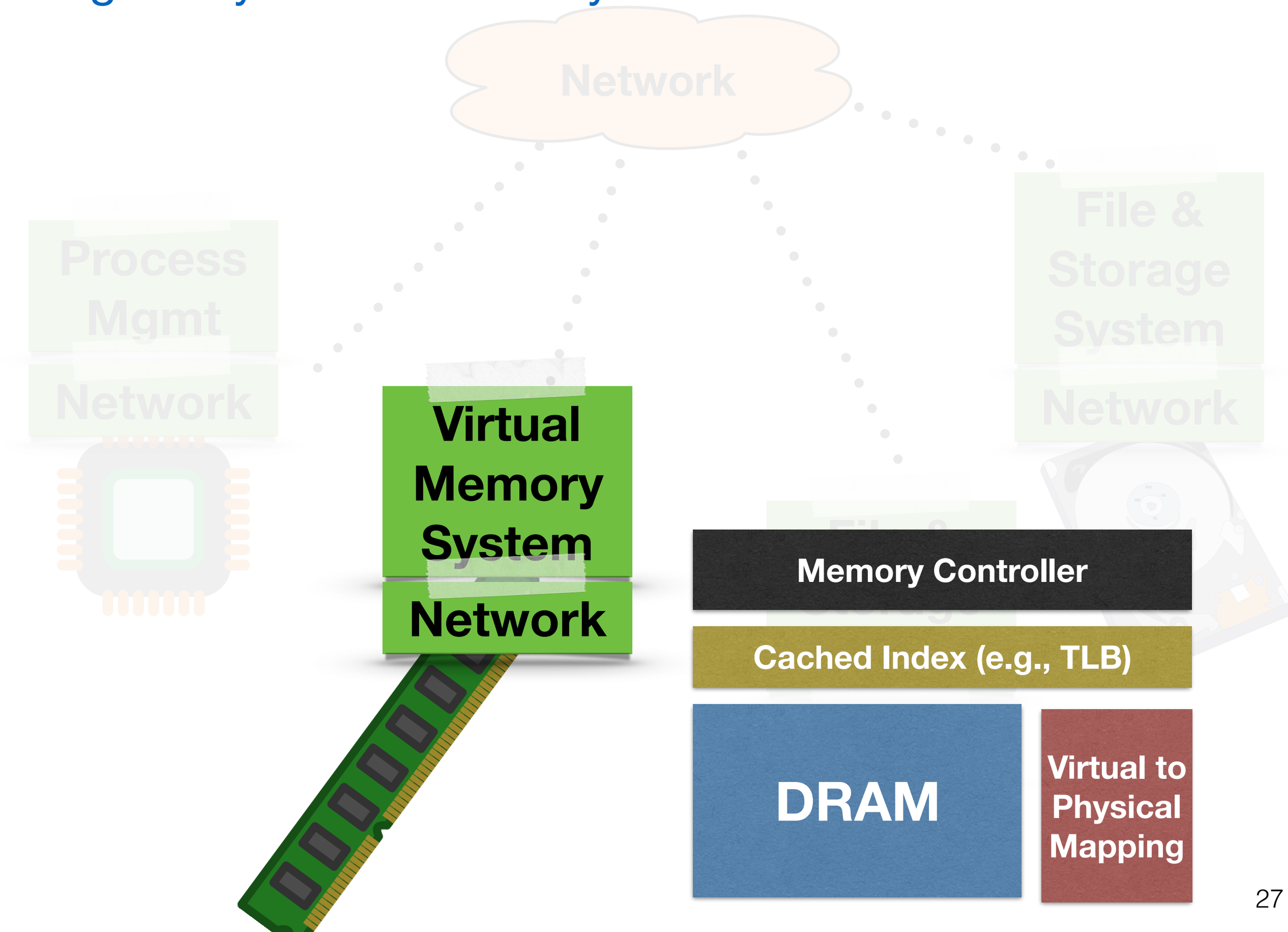
Memory Controller

Cached Index (e.g., TLB)

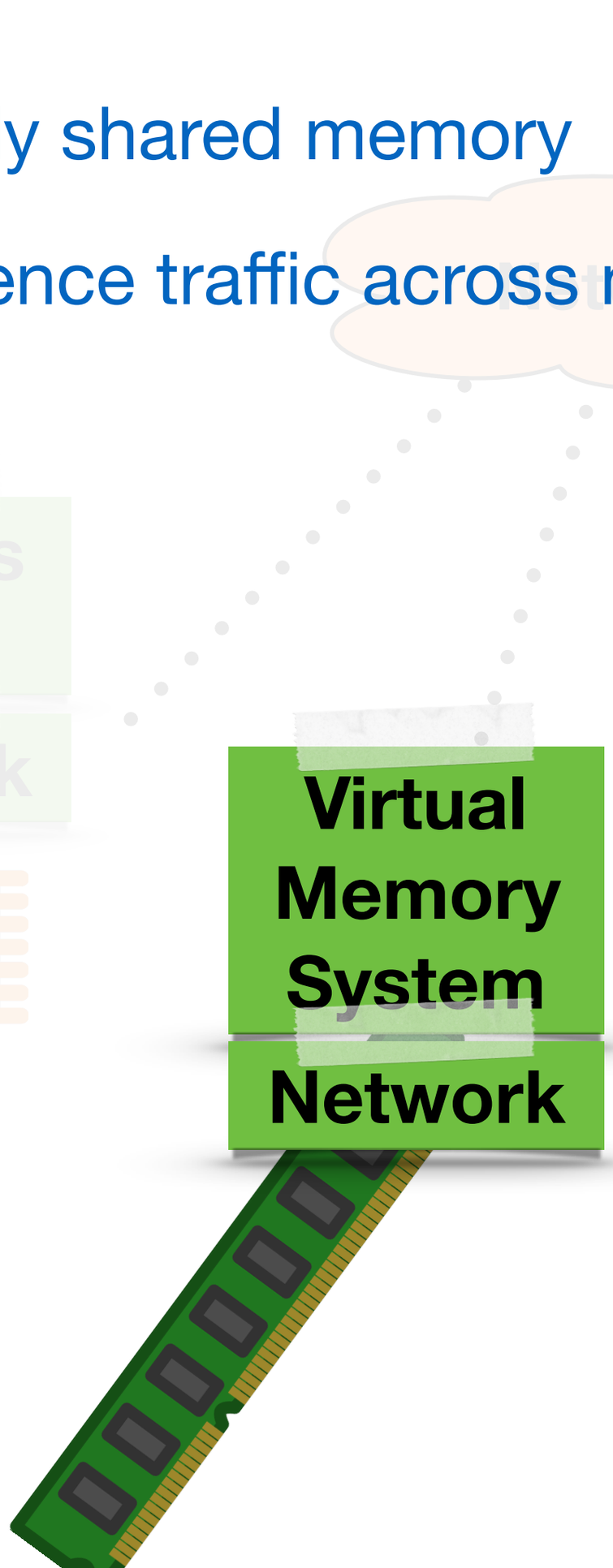
DRAM

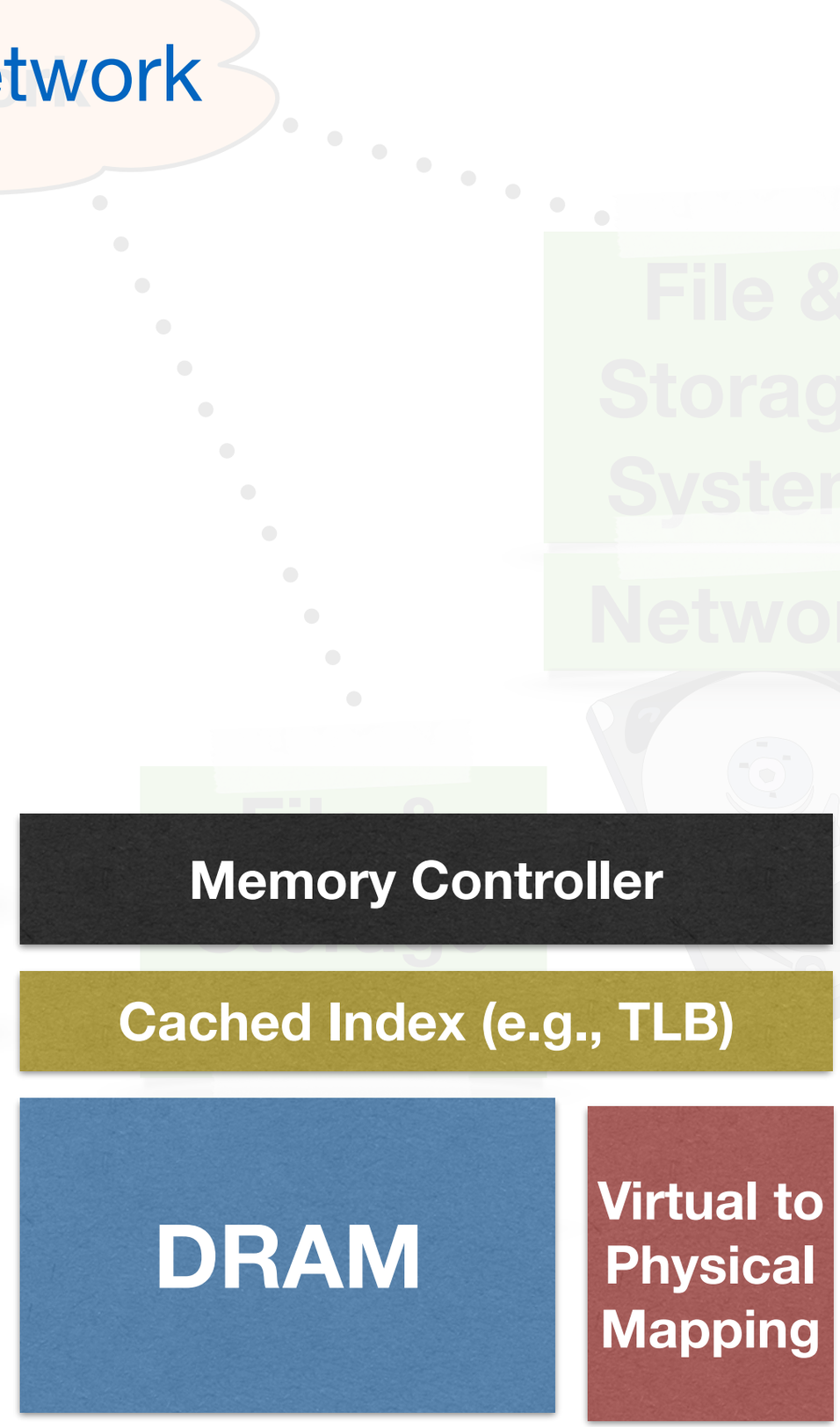
Virtual to Physical Mapping

- No globally shared memory

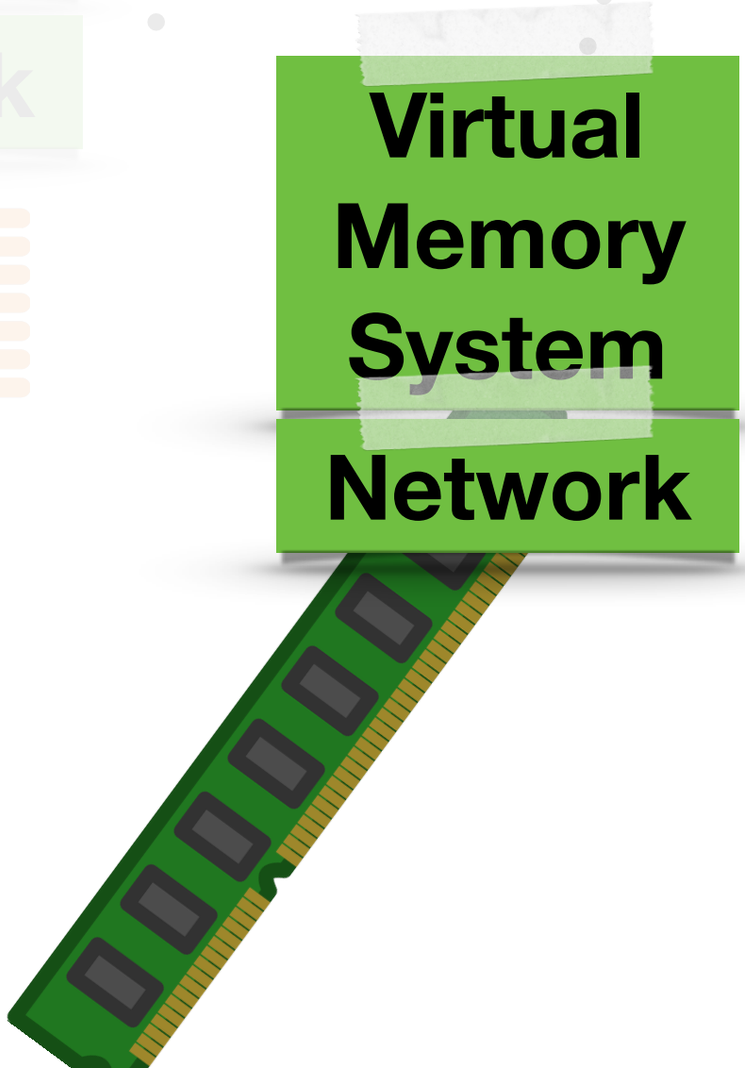


- No globally shared memory
- No coherence traffic across network





- No globally shared memory
- No coherence traffic across network
- RDMA-based RPC



Memory Controller

Cached Index (e.g., TLB)

DRAM

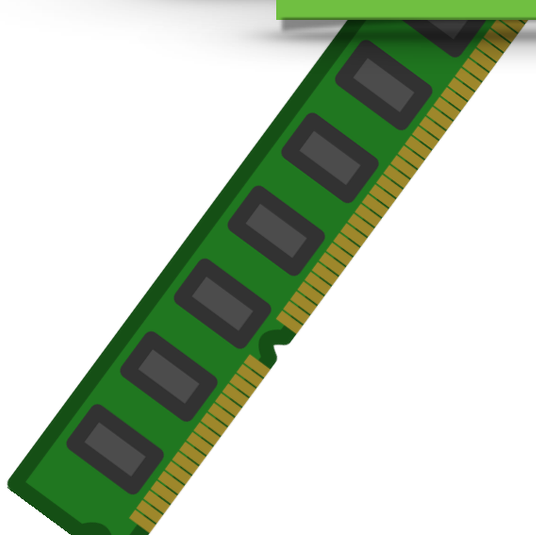
Virtual to Physical Mapping

- No globally shared memory
- No coherence traffic across network
- RDMA-based RPC

Kernel Space
(running in mem controller)

Virtual Memory System

Network



Memory Controller

Cached Index (e.g., TLB)

DRAM

Virtual to Physical Mapping

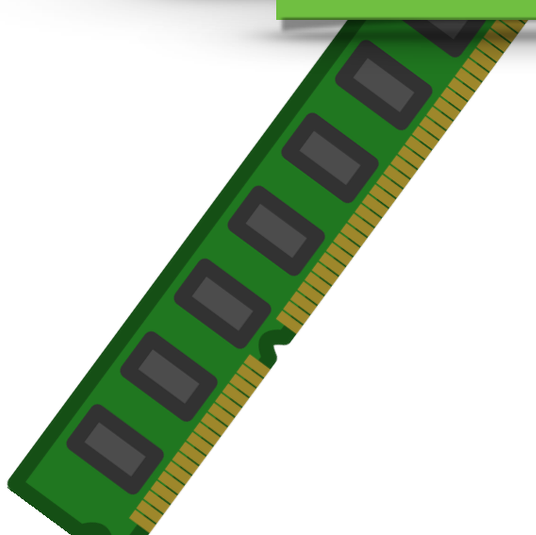
- No globally shared memory
- No coherence traffic across network
- RDMA-based RPC

Virtual address (de)allocation

Kernel Space
(running in mem controller)

Virtual Memory System

Network



Memory Controller

Cached Index (e.g., TLB)

DRAM

Virtual to Physical Mapping

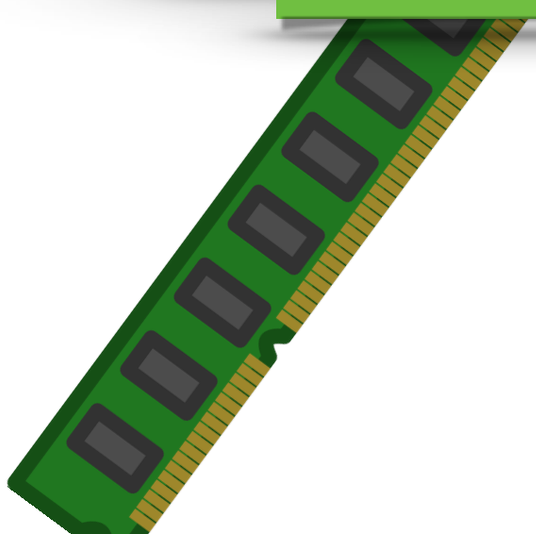
- No globally shared memory
- No coherence traffic across network
- RDMA-based RPC

Kernel Space
(running in mem controller)

Virtual Memory System

Virtual address (de)allocation

Physical address (de)allocation



Memory Controller

Cached Index (e.g., TLB)

DRAM

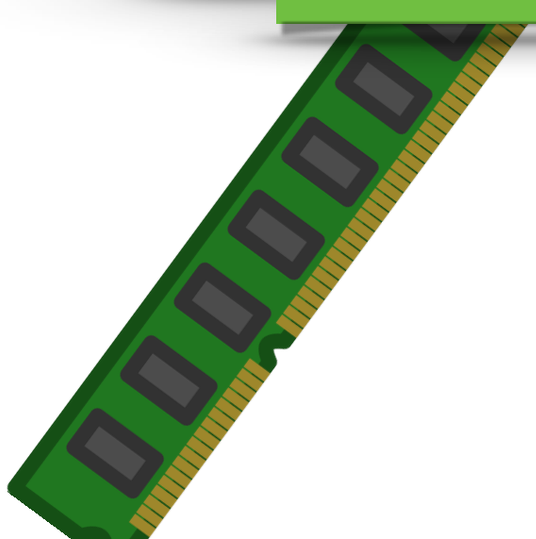
Virtual to Physical Mapping

- No globally shared memory
- No coherence traffic across network
- RDMA-based RPC

Kernel Space
(running in mem controller)

Virtual Memory System

Network



Virtual address (de)allocation

Physical address (de)allocation

Memory-mapped file mgmt

Memory Controller

Cached Index (e.g., TLB)

DRAM

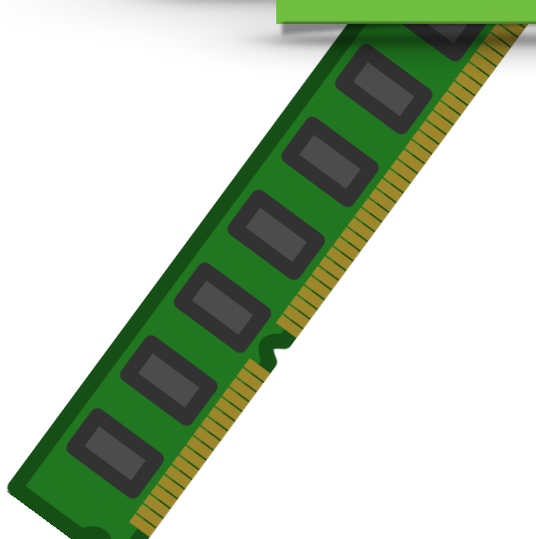
Virtual to Physical Mapping

- No globally shared memory
- No coherence traffic across network
- RDMA-based RPC

Kernel Space
(running in mem controller)

Virtual Memory System

Network



Virtual address (de)allocation

Physical address (de)allocation

Memory-mapped file mgmt

Memory replication

Memory Controller

Cached Index (e.g., TLB)

DRAM

Virtual to Physical Mapping

Challenges

- Cleanly separate OS services
- Fit hardware constraints
- Handle failures
- Global resource management

Status Report

- 170K LOC so far
- Simple processor, memory, storage managers
- Support X86-64
- Backward compatible with common Linux interface
- Run unmodified datacenter application binaries
- Emulate hardware devices using commodity servers

Status Report

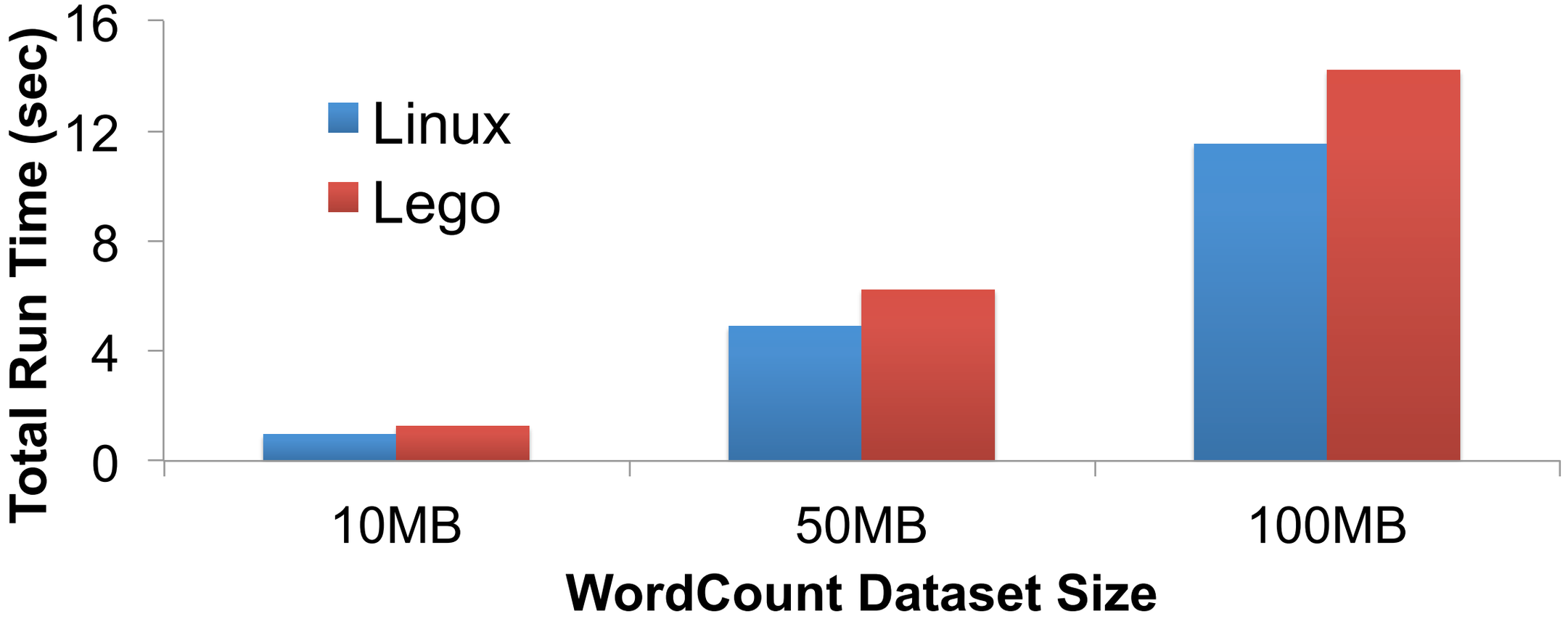
- 170K LOC so far
- Simple processor, memory, storage managers
- Support X86-64
- Backward compatible with common Linux interface
- Run unmodified datacenter application binaries
- Emulate hardware devices using commodity servers

We will open source!

Initial Results are Encouraging

Phoenix (single-node MapReduce), unmodified statically-linked binary

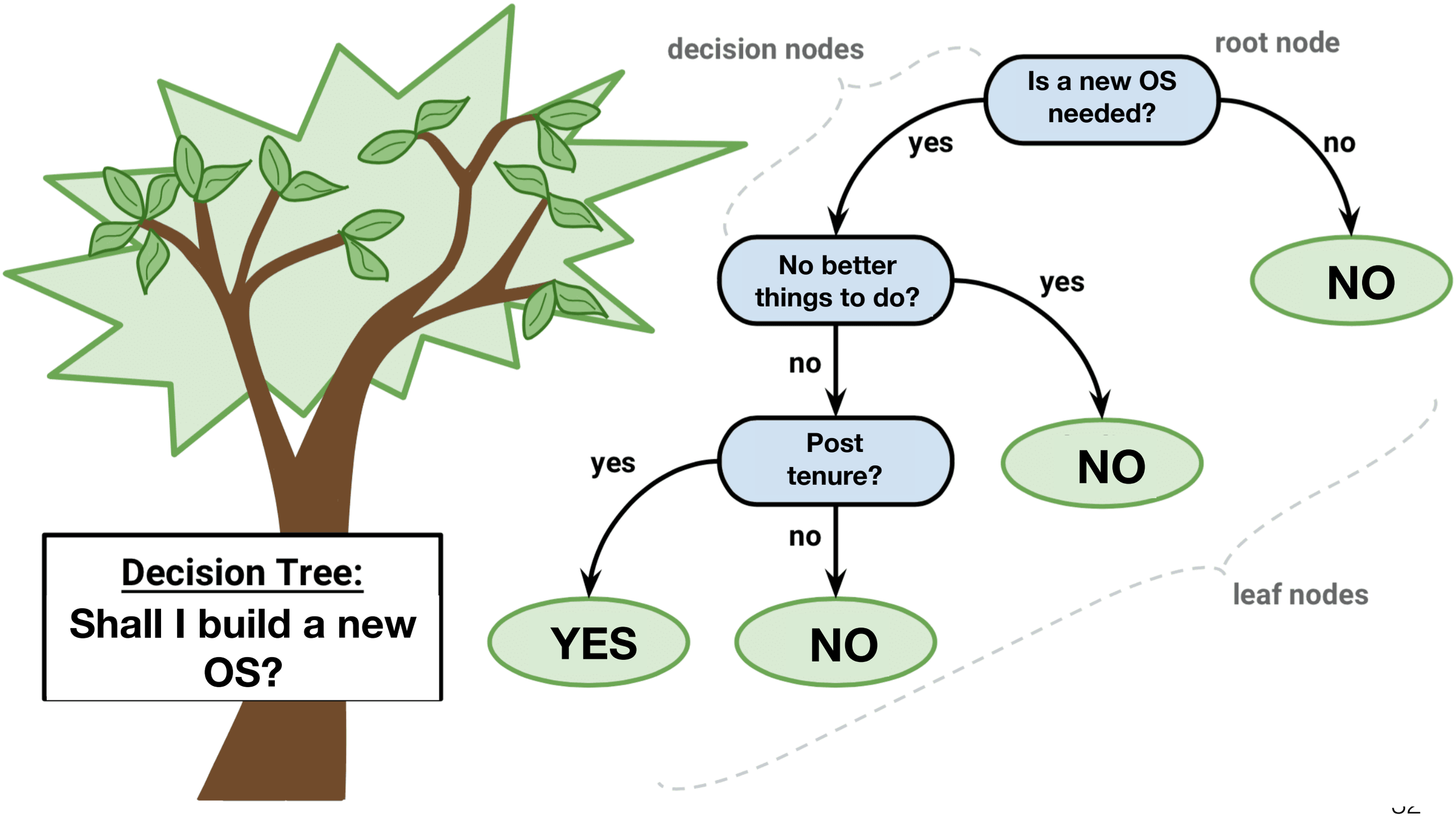
Compare one commodity server running Linux with Lego running on one proc, one mem, one storage, emulated using three servers



Conclusion - A Bunch of Questions

- Time to change datacenters?
- Do you believe in resource disaggregation?
- New OS for new hardware?
- Are we reinventing the wheel?
- Killer applications?

Conclusion [hidden version]



Thank You
Questions?

WukLab

wuklab.io

PURDUE UNIVERSITY