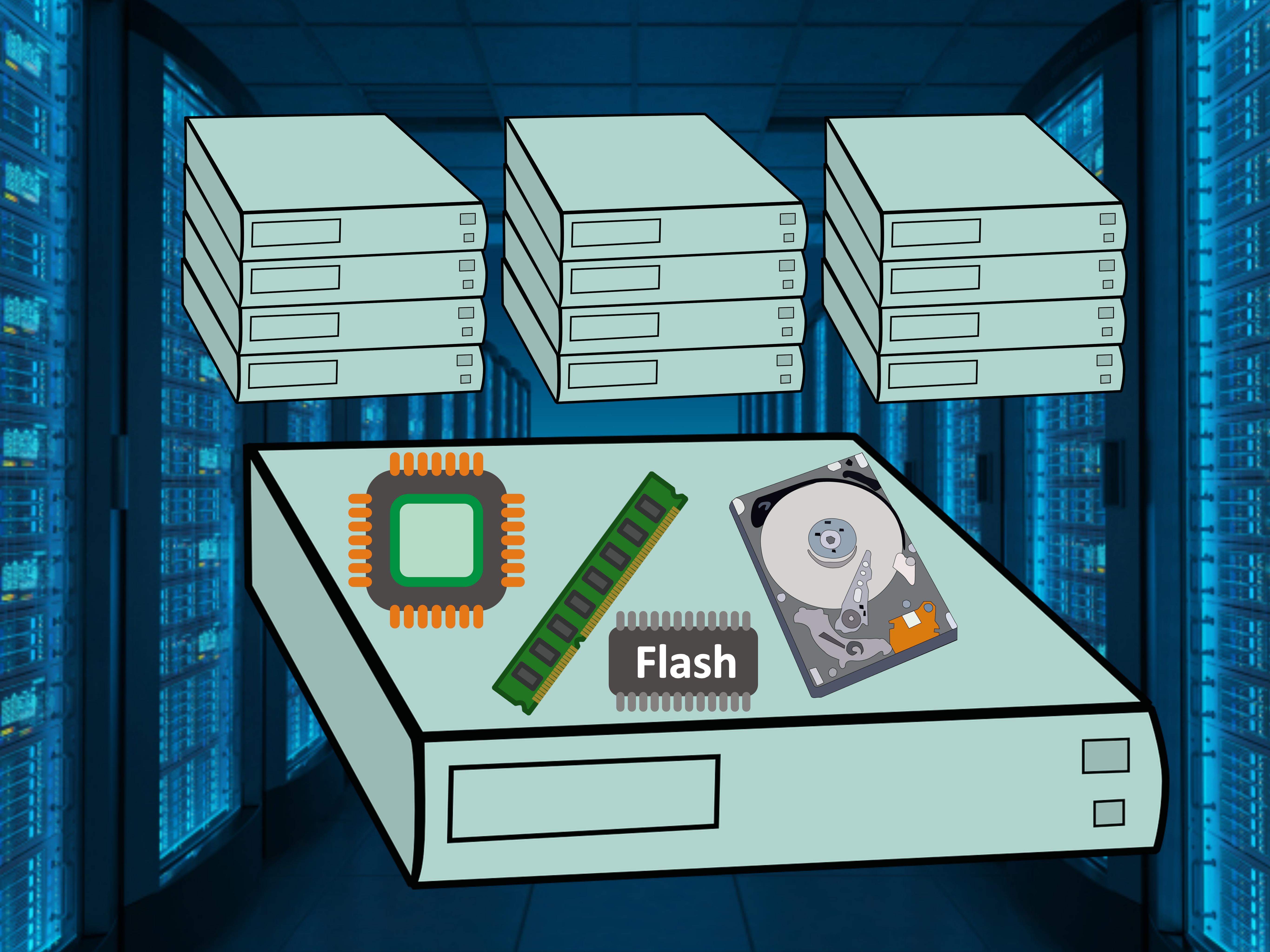
Disaggregated Operating System

Yiying Zhang

Yizhou Shan, Yilun Chen, Yutong Huang, Sumukh Hallymysore

WukLab

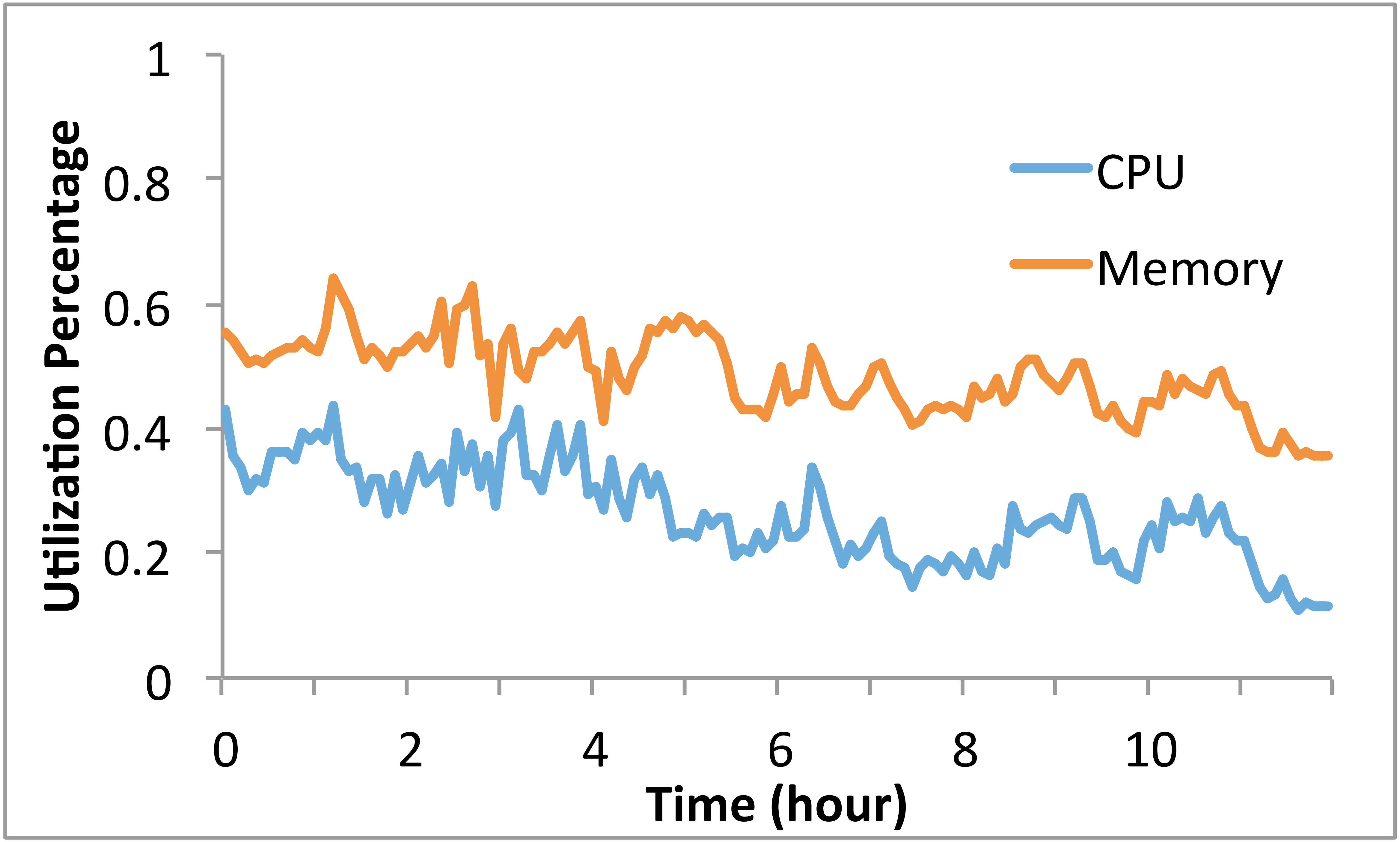
PURDUE UNIVERSITY


Flash

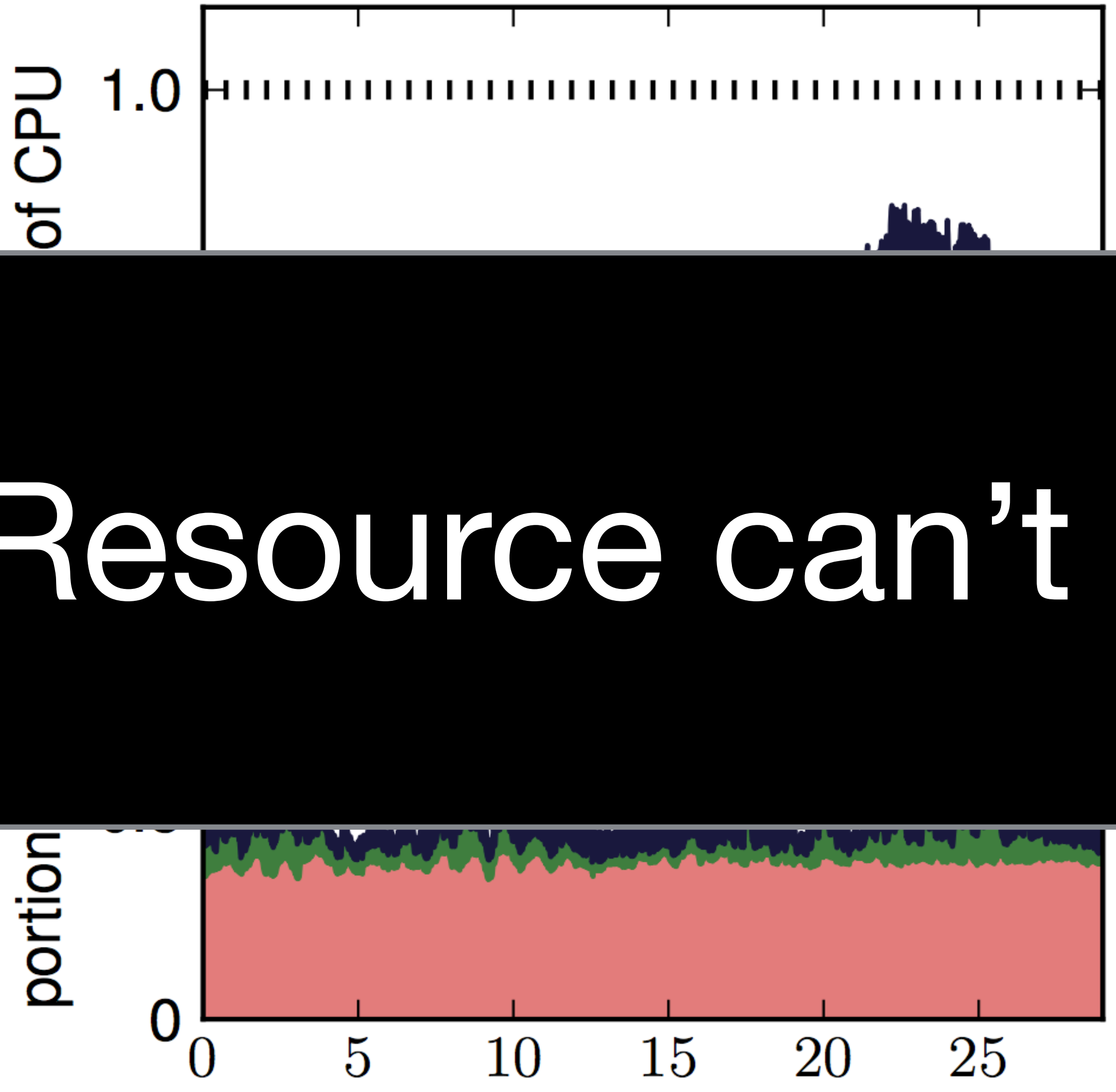
Monolithic Server

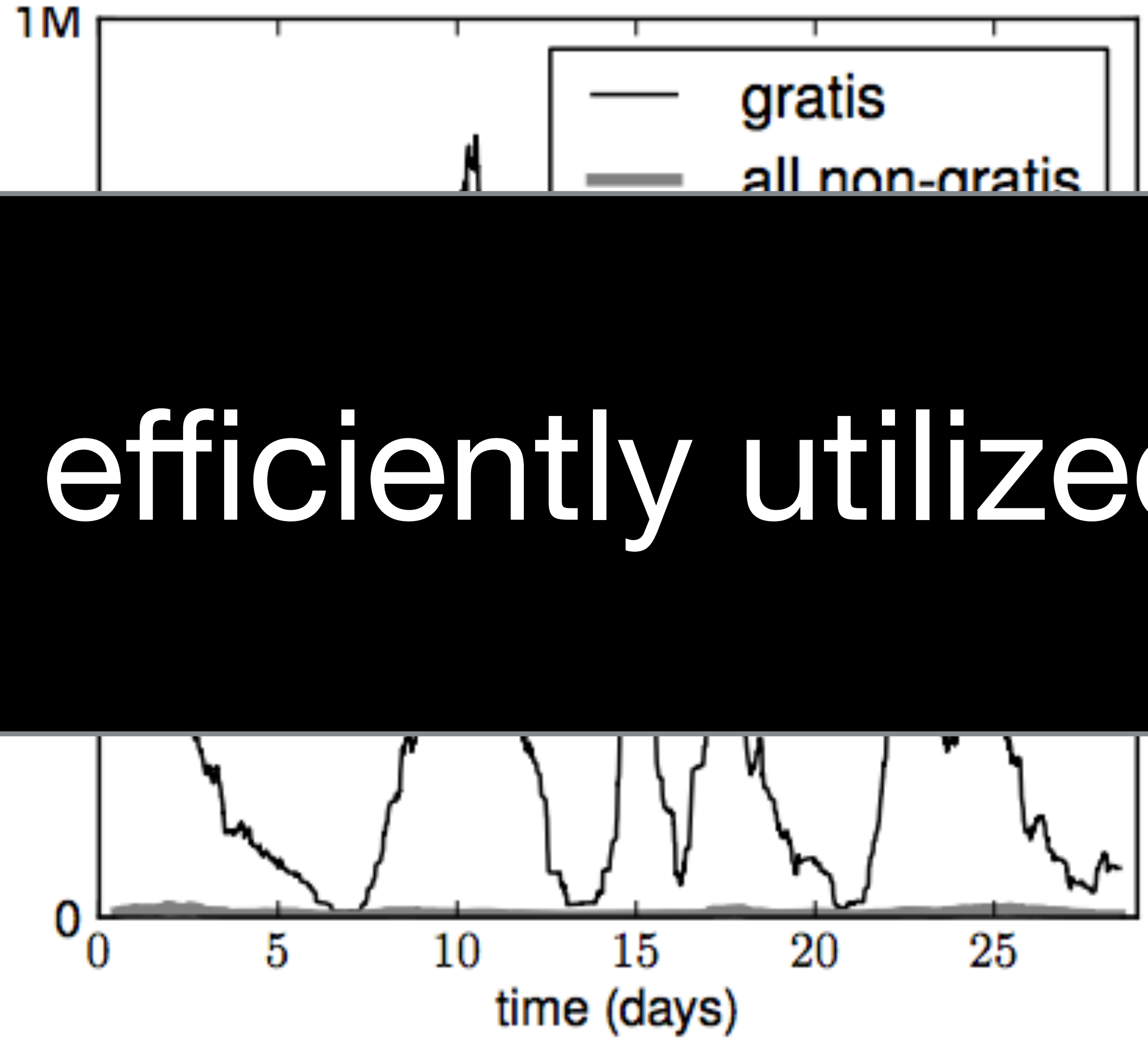
- Resource utilization
- Failure
- Flexibility

Alibaba Cluster Resource Utilization



Google Cluster Resource Utilization

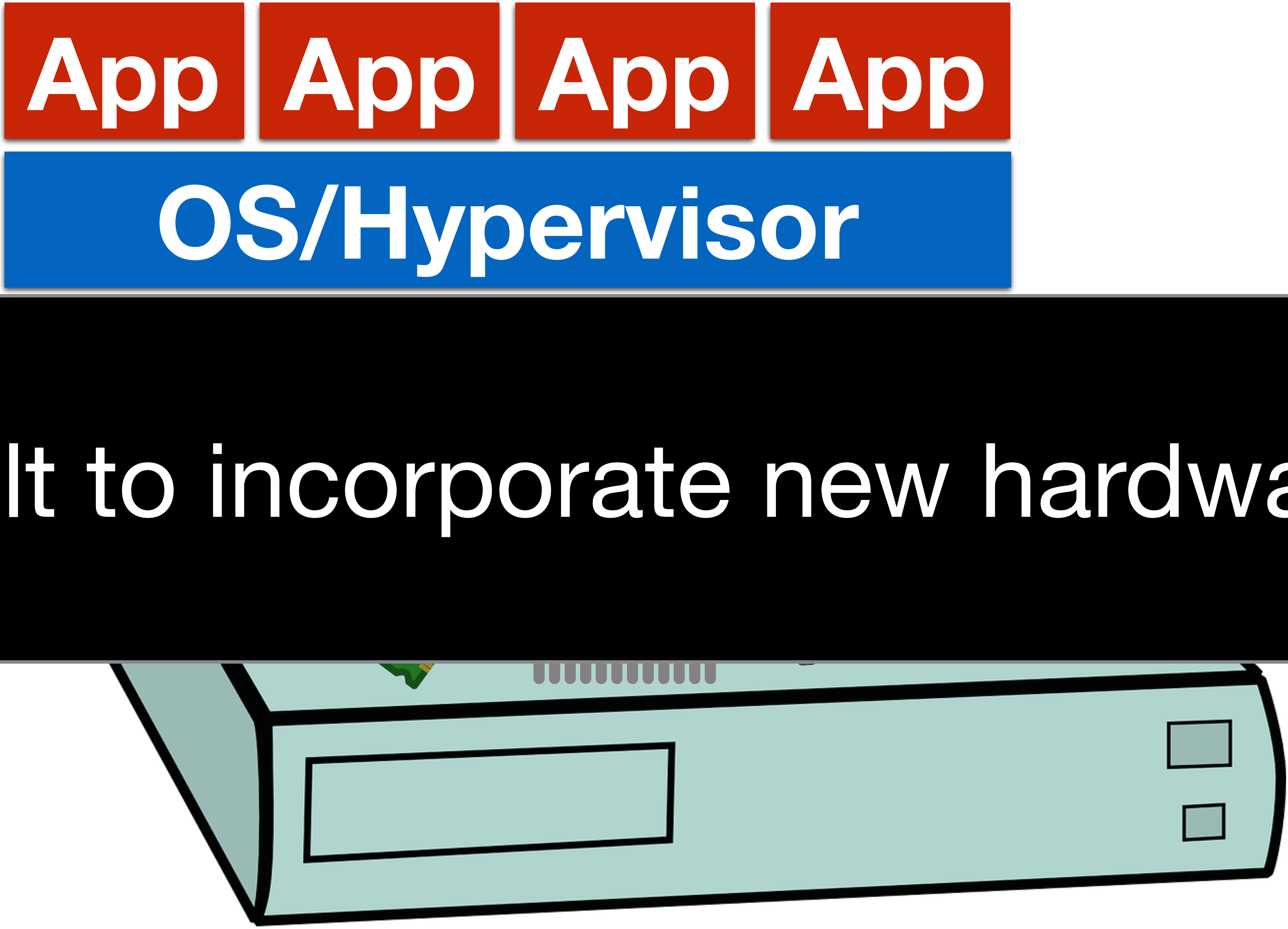




Resource can’t be efficiently utilized

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

No fine-grained failure handling



Difficult to incorporate new hardware

Monolithic Server

- Resource utilization
- Failure
- Flexibility
- Memory capacity wall

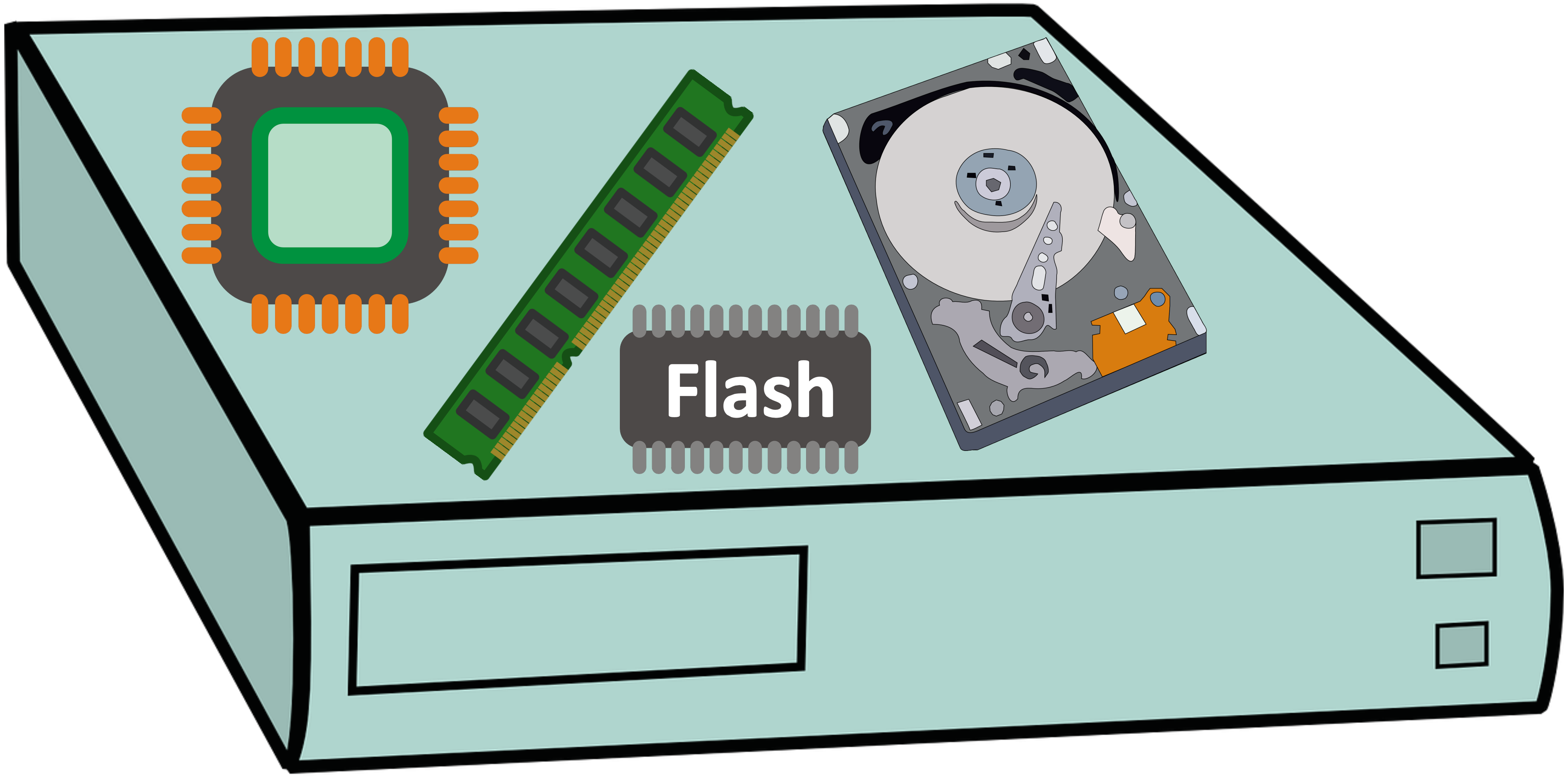
When was the last time

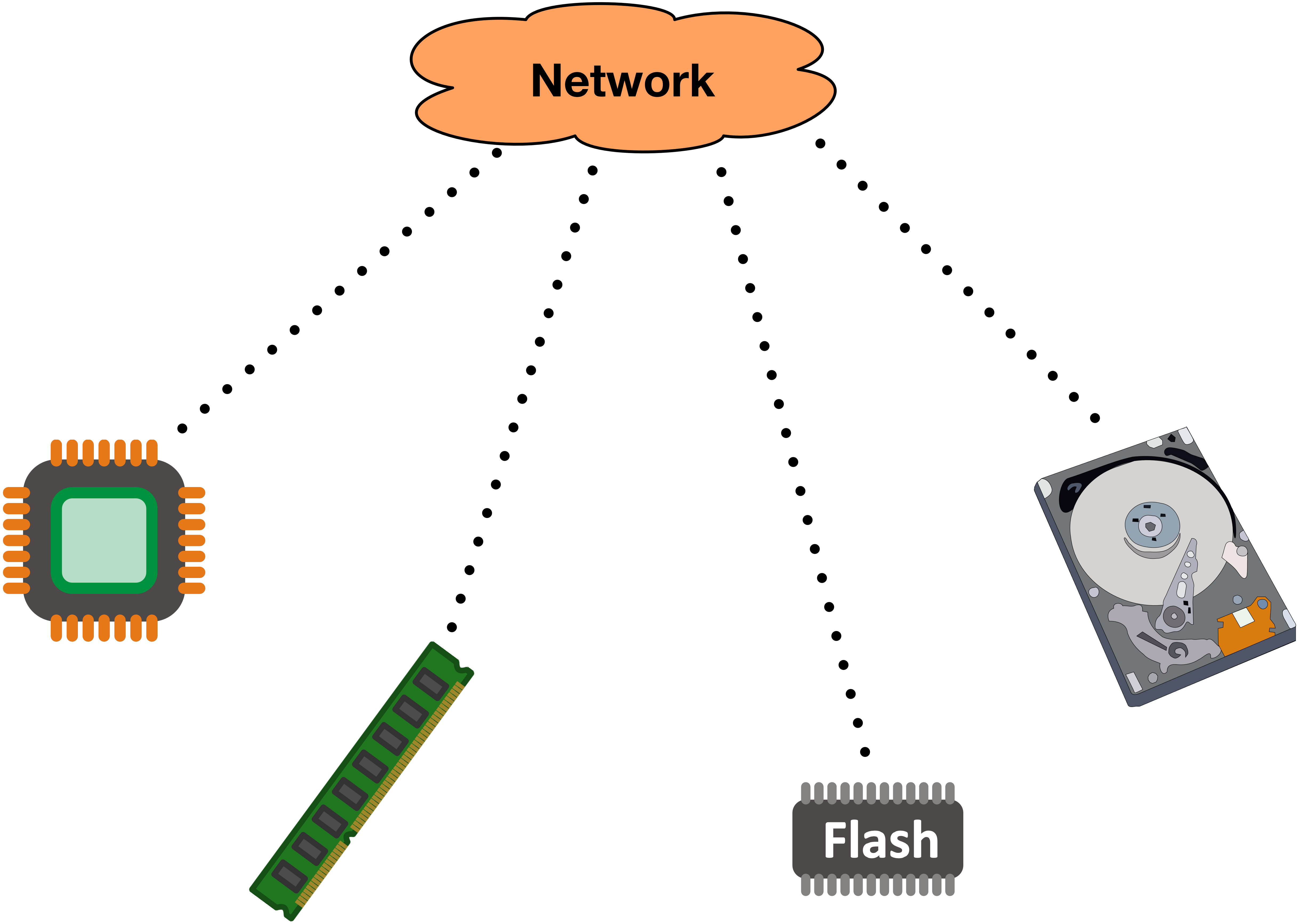
you did something for the first time?

TIME FOR CHANGE

Resource Disaggregation:

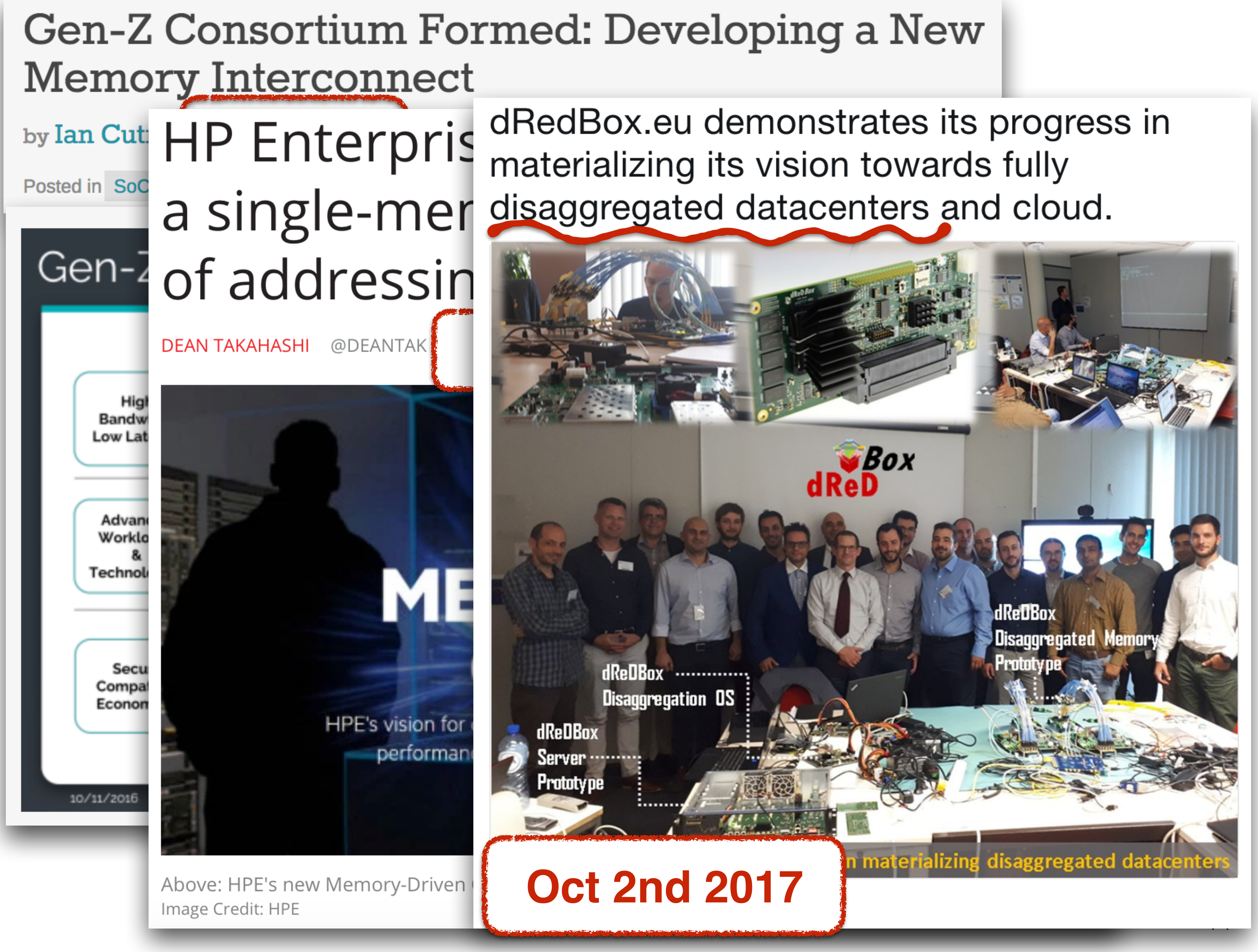
Breaking monolithic servers into network-attached, independent hardware components


Flash



Network

Flash



Gen-Z Consortium Formed: Developing a New Memory Interconnect

by Ian Cut

Posted in SoC

HP Enterpris a single-mer of addressin

DEAN TAKAHASHI @DEANTAK

Above: HPE's new Memory-Driven

Image Credit: HPE

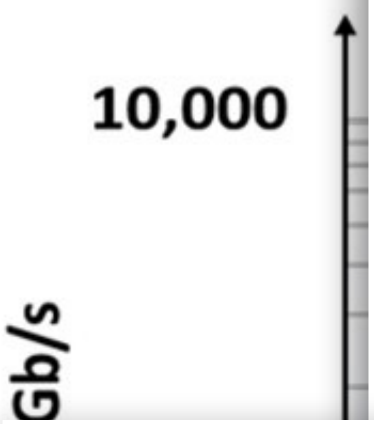
dRedBox.eu demonstrates its progress in materializing its vision towards fully disaggregated datacenters and cloud.

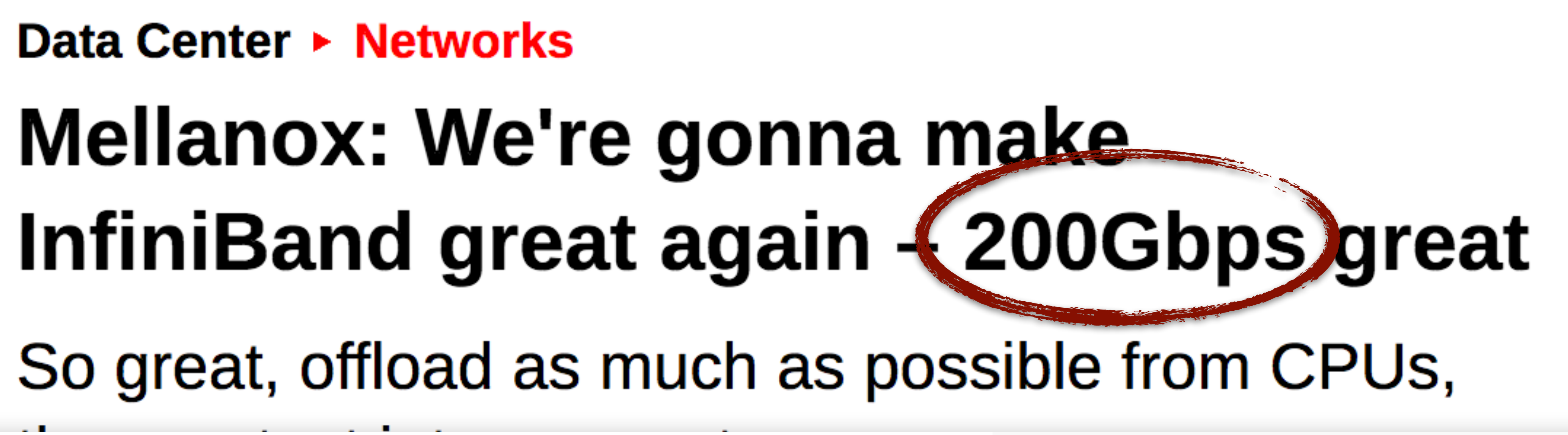
dReD Box
dReDBox Disaggregated Memory Prototype
dReDBox Disaggregation OS
dReDBox Server Prototype
n materializing disaggregated datacenters

Oct 2nd 2017

Why Now?

- Faster network

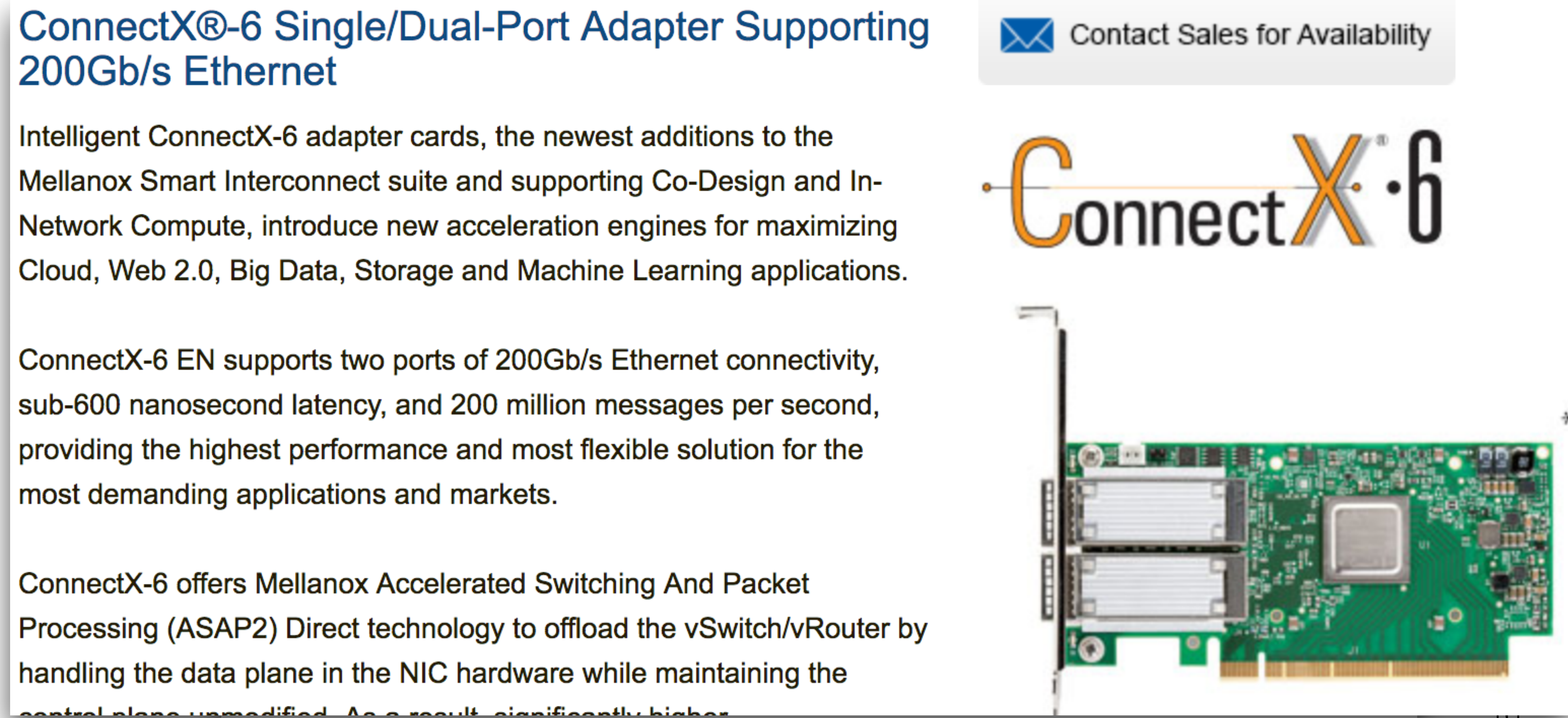




Data Center ▸ Networks

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

So great, offload as much as possible from CPUs,



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

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 control plane unmodified. As a result, significantly higher

90ns and aggregate capacity is 16Tbps.

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

Using Existing Kernels?

- Monolithic/micro kernel: built for single monolithic server

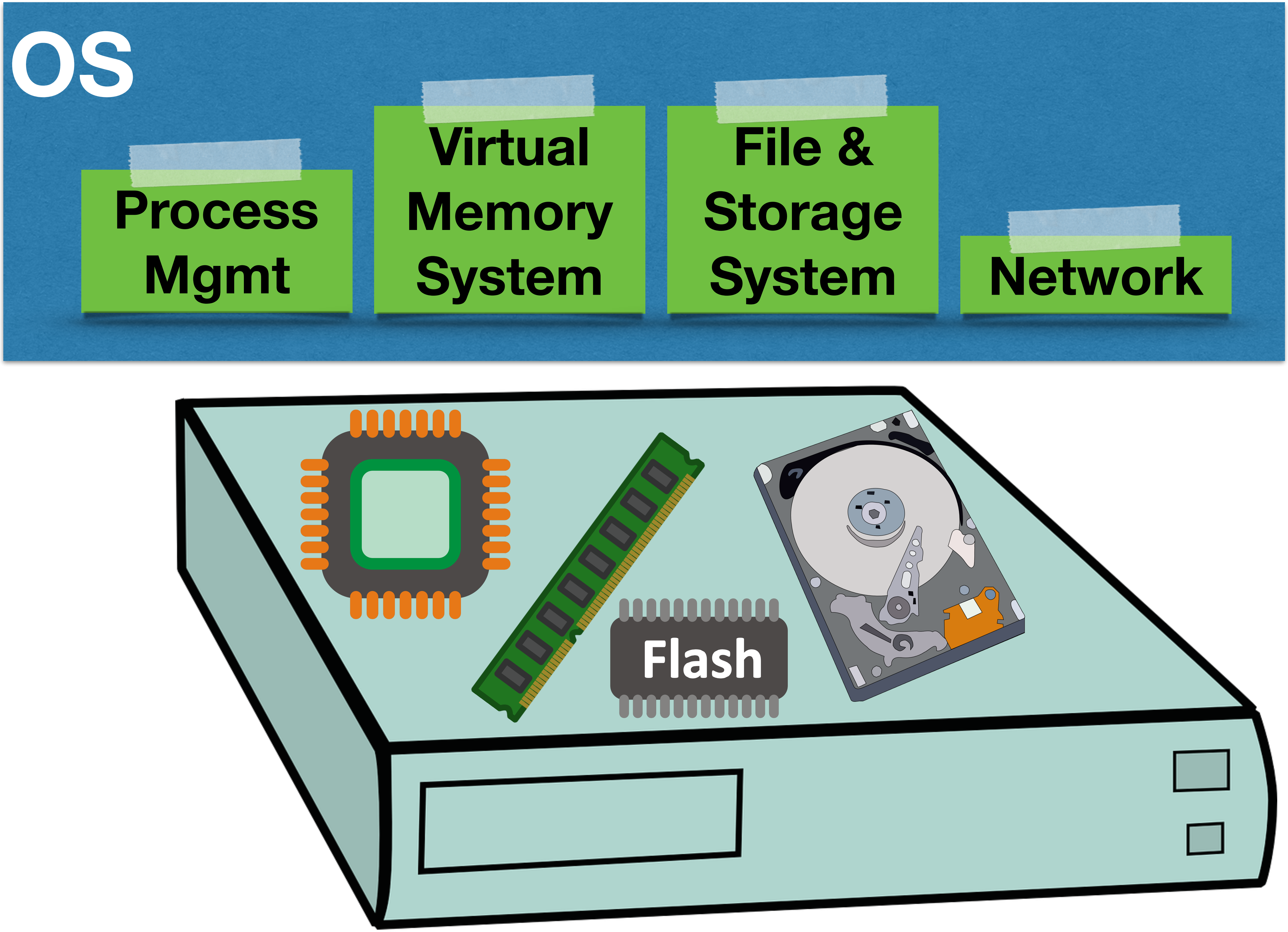
- Multikernel: (vertically) replicated kernel across cores

- 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!



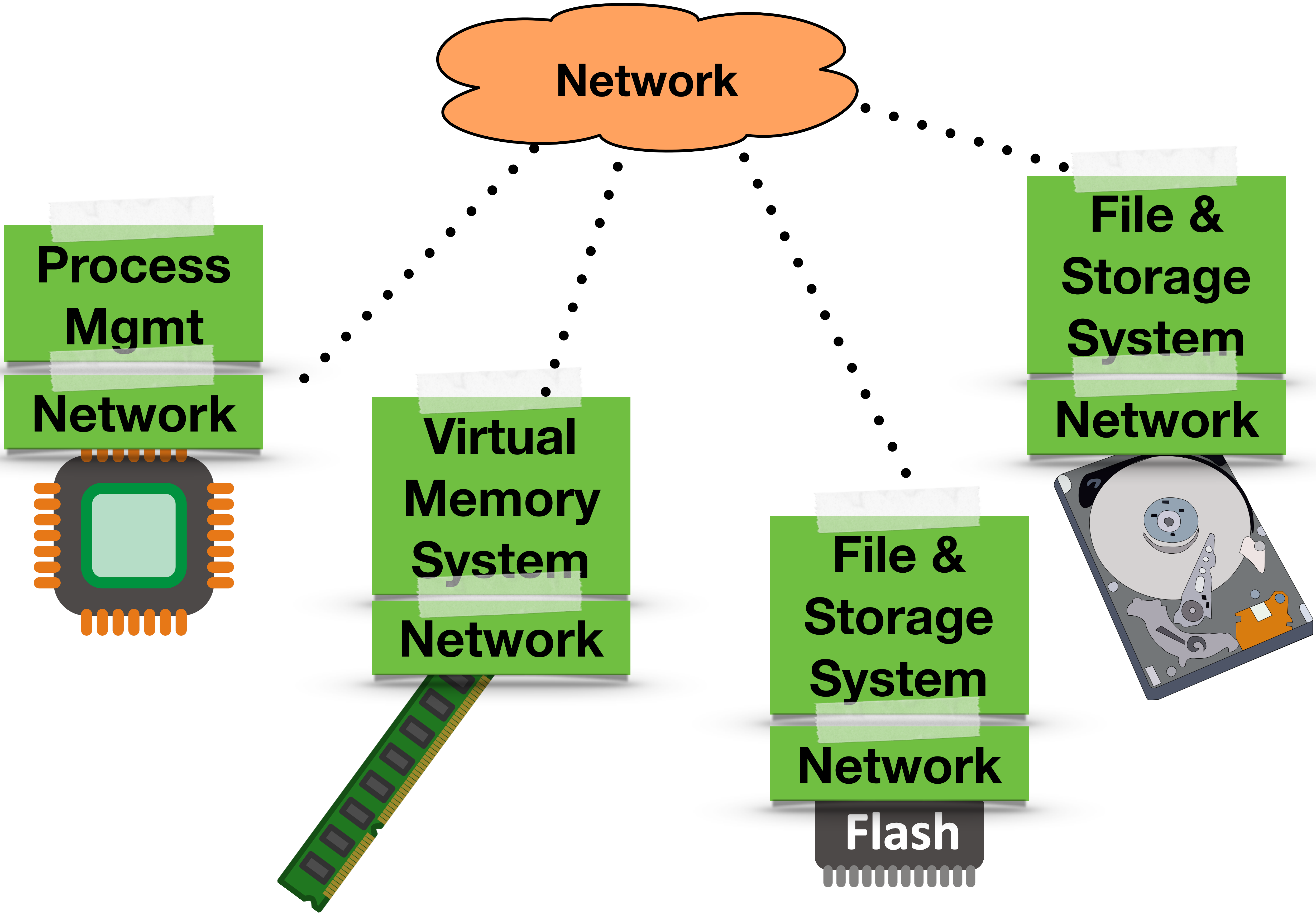
OS

Process Mgmt

Virtual Memory System

File & Storage System

Network



Network

Process Mgmt

Network

Virtual Memory System

Network

File & Storage System

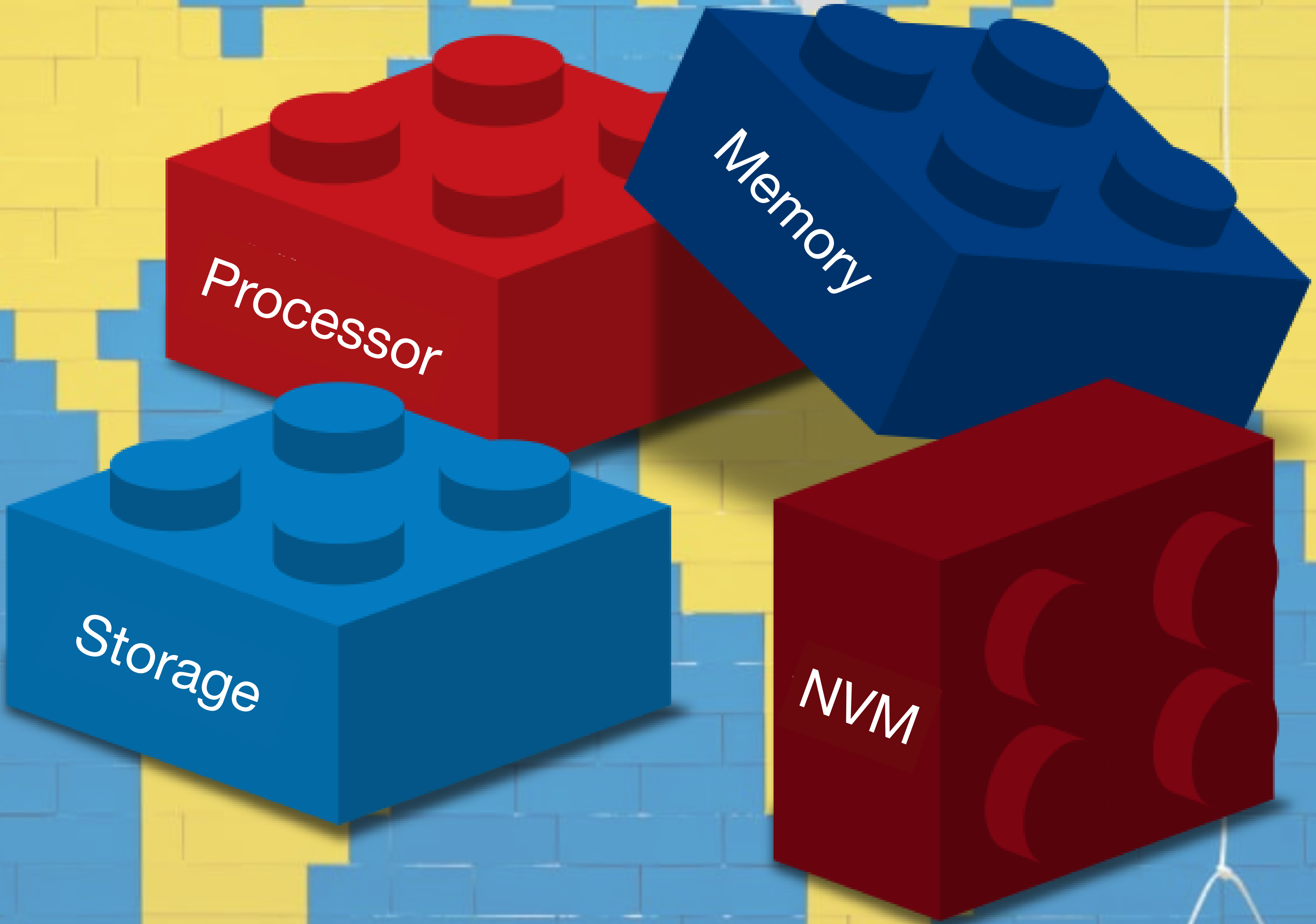
Network

Flash

File & Storage System

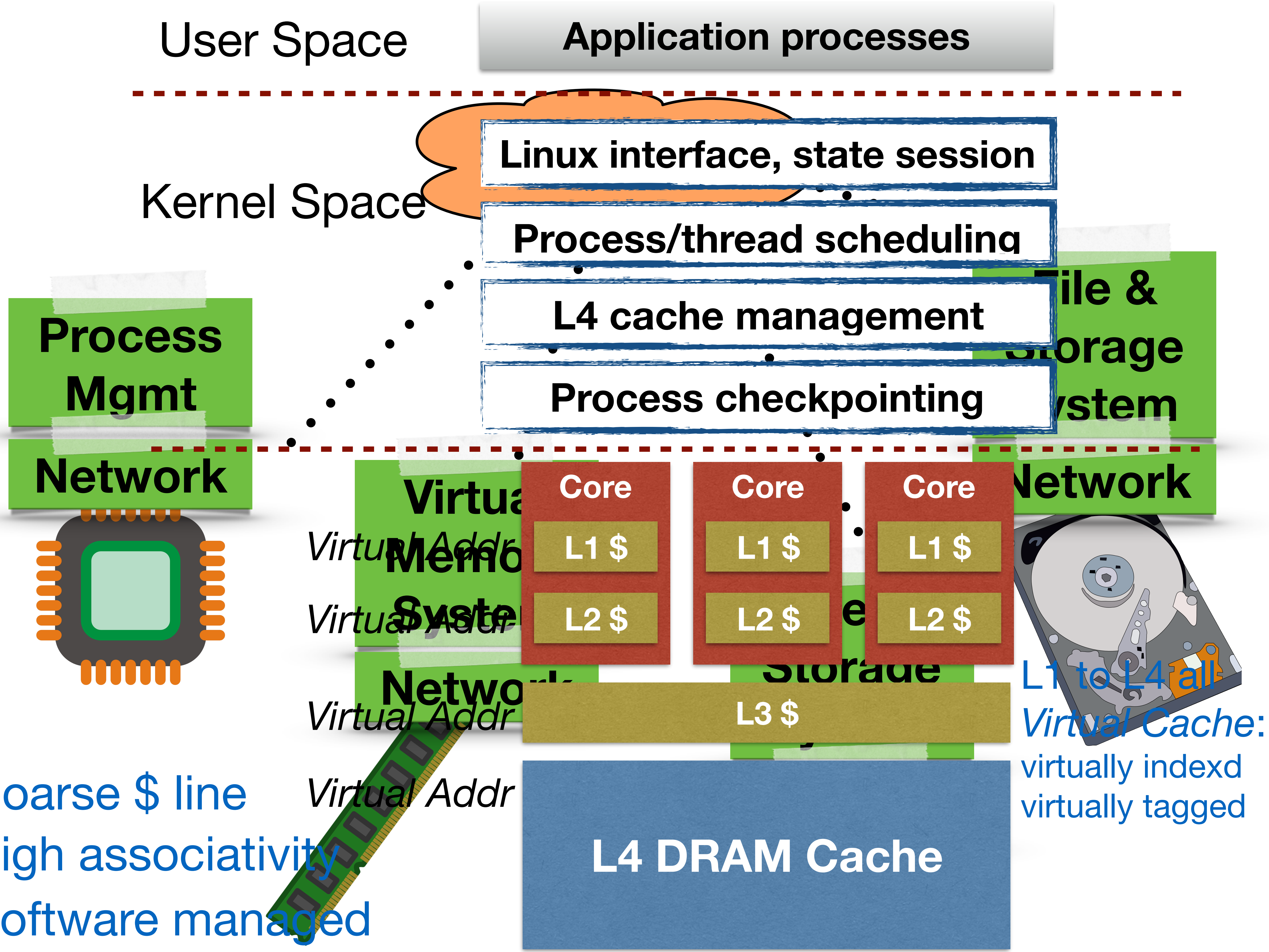
Network

Lego



Challenges

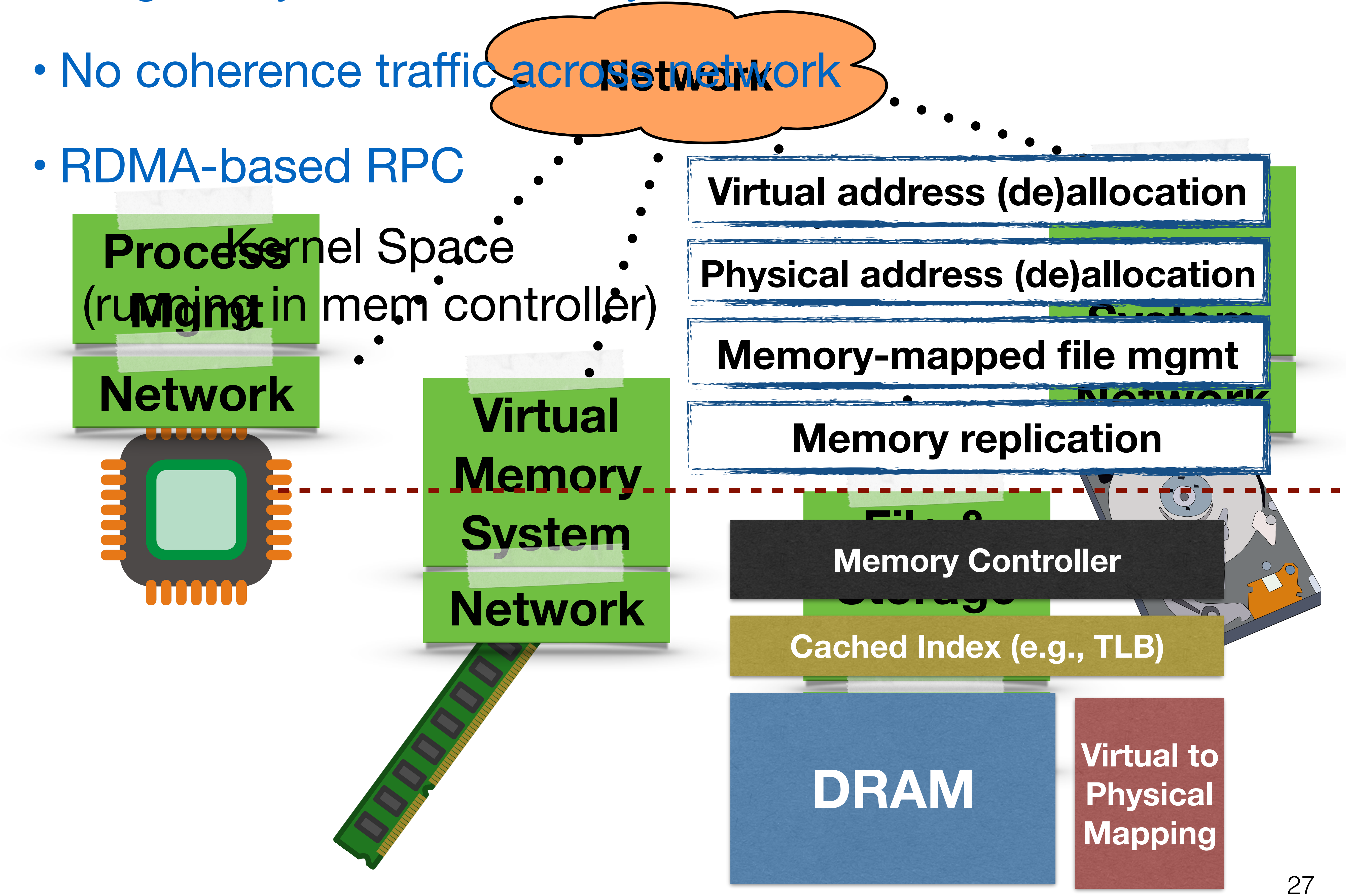
- Cleanly separate OS services
  - *Stateless, minimal dependencies*
- Fit hardware constraints
  - Processor: no or limited local DRAM
  - Memory: limited processing power



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

- Coarse $ line
- High associativity
- Software managed

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



Network

Process Mgmt

Network

Kernel Space (running in mem controller)

Virtual Memory System

Network

Virtual address (de)allocation

Physical address (de)allocation

Memory-mapped file mgmt

Memory replication

System

Network

File & Storage

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

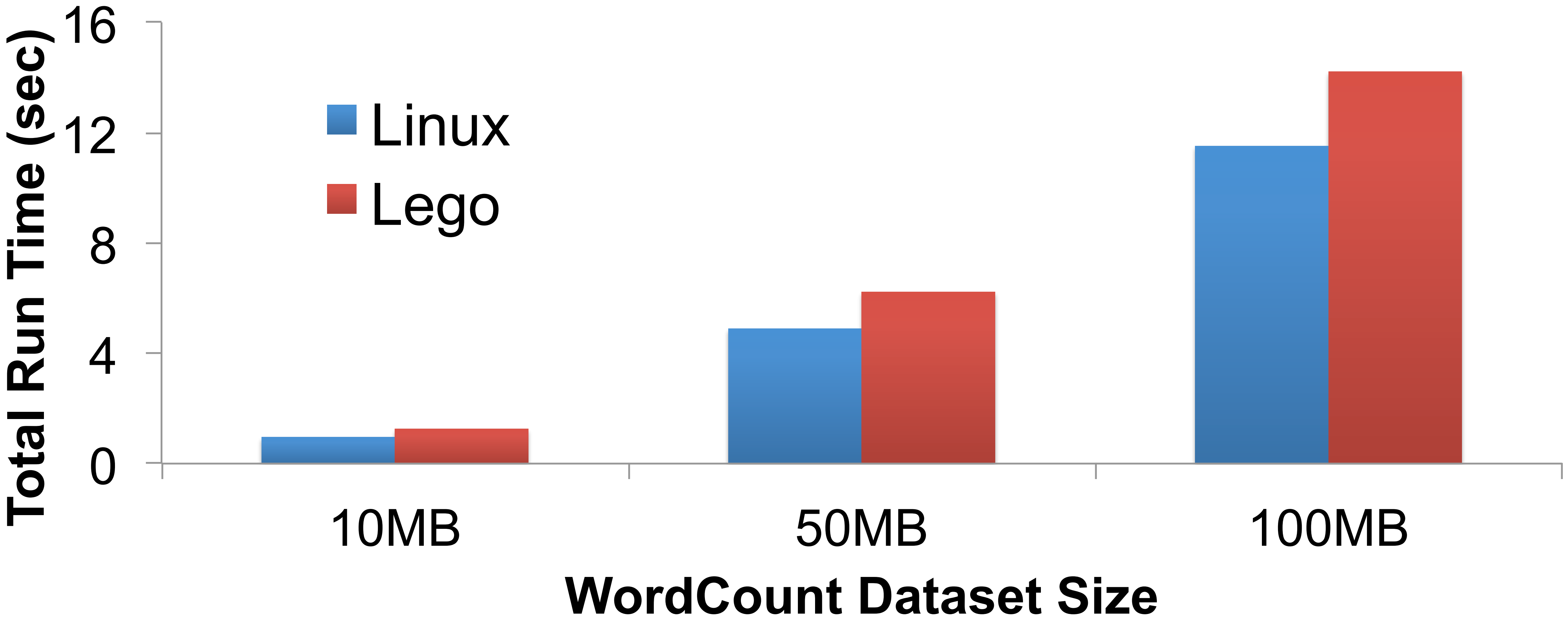
- 170K LOC so far
- Simple processor, memory, storage managers
- Support X86-64
- Backward compatible with common Linux interface
- Run unmodified datacenter application binaries
- Emulate hardware devices using commodity servers

We will open source!

Initial Results are Encouraging

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

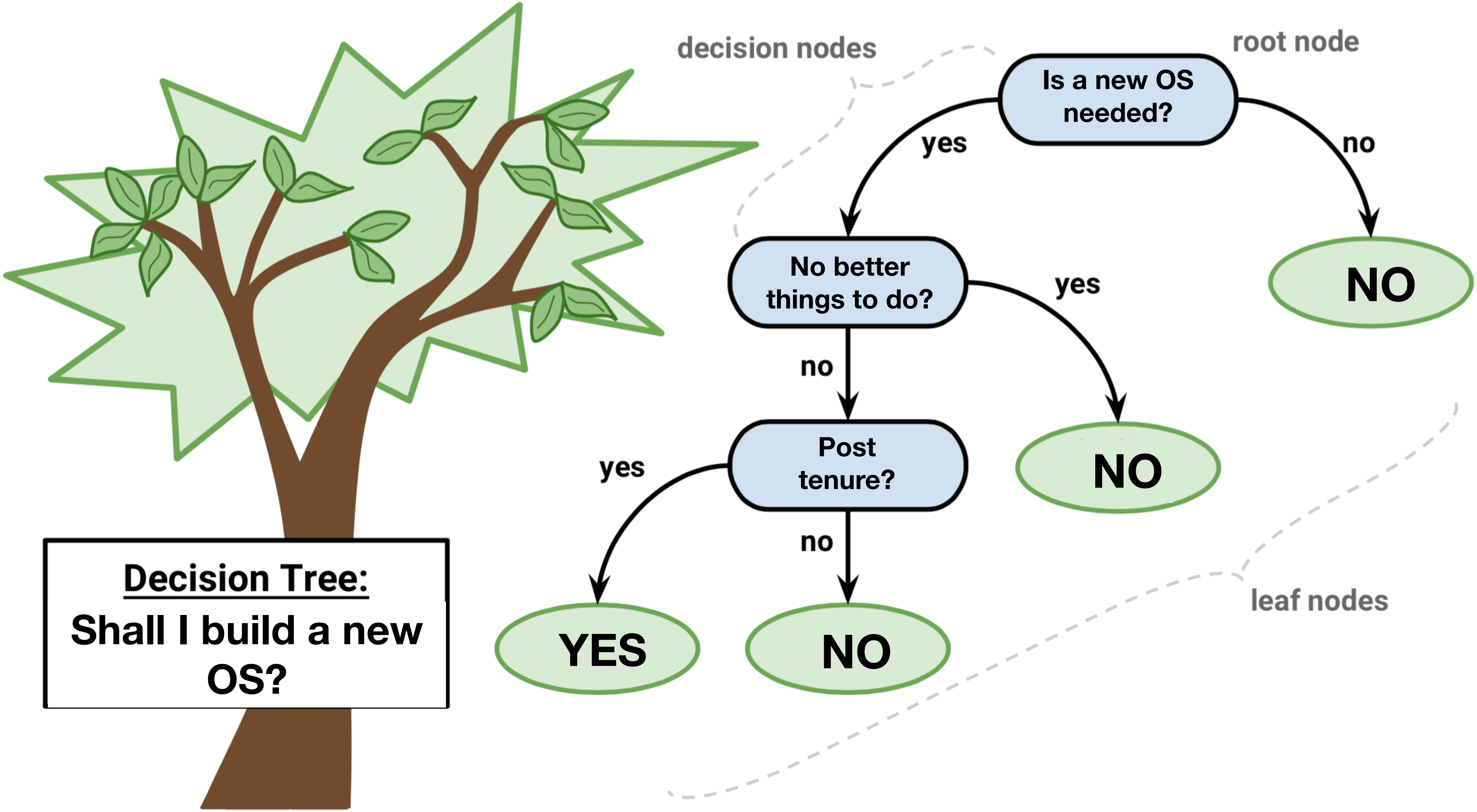
Compare one commodity server running Linux with Lego running on one proc, one mem, one storage, emulated using three servers



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]



Decision Tree:

Shall I build a new OS?

Thank You
Questions?

WukLab

wuklab.io

PURDUE
UNIVERSITY