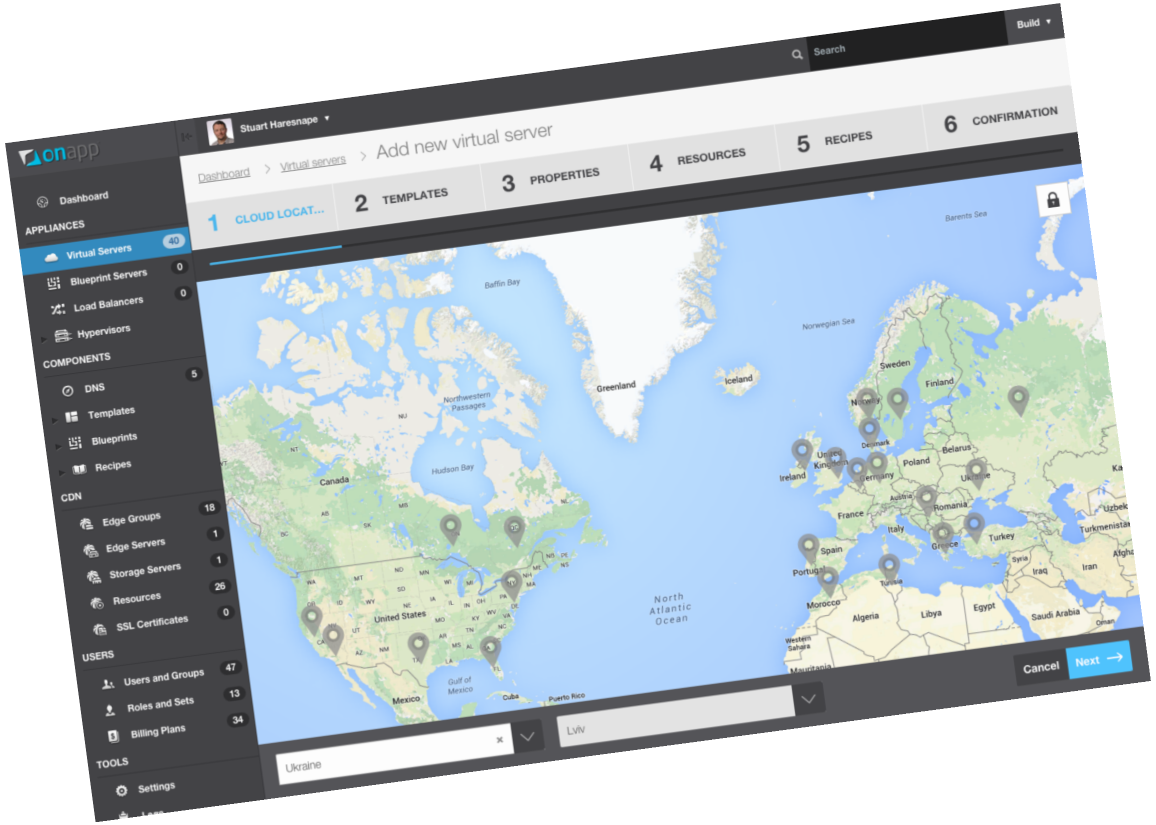
Hypervisor support for emerging scale-out / scale-up architectures

Julian Chesterfield
Chief Scientific Officer, OnApp Ltd
julian@onapp.com

Brief Intro to OnApp

- Company founded **2010**
- Spun out of a major service provider following acquisition by **Lloyds Bank**
- 180 full time employees, **HQ in london**
- Offices on 3 continents
- OnApp powers **1 in 3** public clouds
    - 4000+ DC/cloud operators



Why is OnApp interested in MicroServers?

- OnApp focus is on next generation scale in public cloud and data centre orchestration
- ***Core density*** and ***power efficiency*** are the top concerns for public cloud operators
- Performance and scalability of storage and network services are a requirement
- ARM-based servers are gaining traction in the DC
- Programmable accelerated IO interfaces are becoming mainstream
  - Hyper-converged Infrastructure (HCI) with accelerated IO
  - Securing tenant workloads in the cloud with hardware assisted encryption of storage and network traffic

Brief explanation of HCI

- Hyper-Converged Infrastructure:
  - Software Defined Compute (Hypervisor Virtualisation)
  - Software Defined Networking (SDN, Openflow etc..)
  - Software Defined Storage (SDS)
- Fastest growing infrastructure orchestration trend in enterprise DC
- SDS - Utilising commodity direct attached storage devices
  - Software controlled distributed block storage for Virtual machines
- Software control is extremely advantageous
  - fast dynamic reconfiguration
  - feature updates
  - no hardware appliance dependency
- But **performance** is significantly impacted

NUTANIX.

Hewlett Packard
Enterprise

simplivity

Web scale computing trends

- Greater Power efficiency demand is driving integrated SoC processor adoption
    - Intel XeonD family
    - Increasing core count, no dependency on NUMA
    - ‘Yosemite’-style architecture with centralised IO resources across SoC nodes
- Dark silicon limitation is generating much greater focus on FPGA and CPU co-processors
- Wide scale adoption of flash storage (up to 16 GBit/s per drive) coupled with high performance ethernet (40/50/100 GBit/s) is driving hardware assisted network storage access (NVMe over Fabric)

ACTICLOUD project - Combatting Resource Under-use in Cloud DCs

- Resource silo units are constrained by the ‘PC’ architecture
  - All cores and memory are coherent
  - Server admins must reserve headroom on each unit for bursts
  - Servers are mirrored for redundancy so the issue is multiplied
- Resource silo units present challenges in efficiently utilising memory
  - Maximum memory for any single VM is constrained by the physical server
  - Server admins typically over-equip servers with costly and energy inefficient memory as a result
  - Bin packing VMs efficiently across the numerous nodes is hard to do efficiently

ACTiCLOUD info

EU H2020 project, Grant Agreement Nº: 732366

Start date: 1 Jan 2017

Duration: 36 months

Partners:

ICCS ΕΠΙΣΕΥ

NUMASCALE

MANCHESTER 1824
The University of Manchester

KALEAO

monetdb solutions

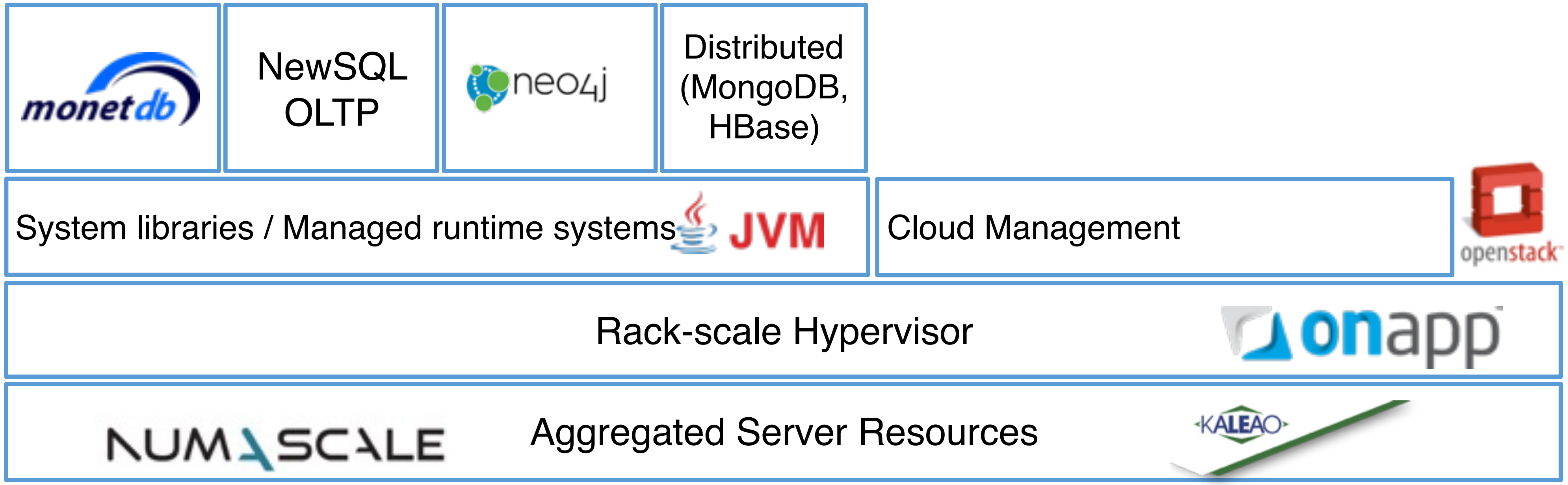
neo4j

onapp

UMEÅ UNIVERSITET

Coordinator: ICCS

Architecture



NewSQL OLTP

Distributed (MongoDB, HBase)

System libraries / Managed runtime systems

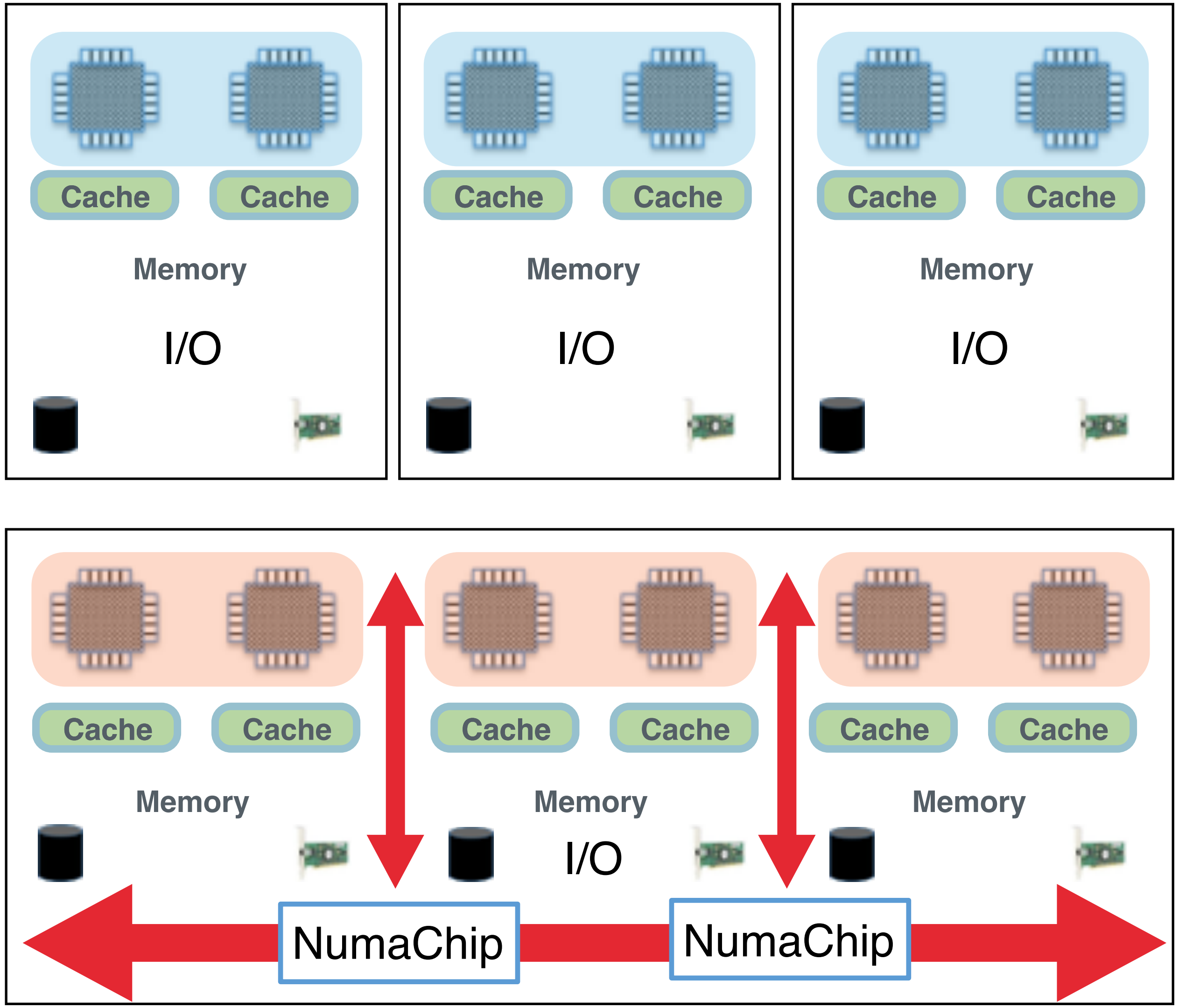
Cloud Management

Rack-scale Hypervisor

Aggregated Server Resources

ACTICLOUD Hardware Architectures

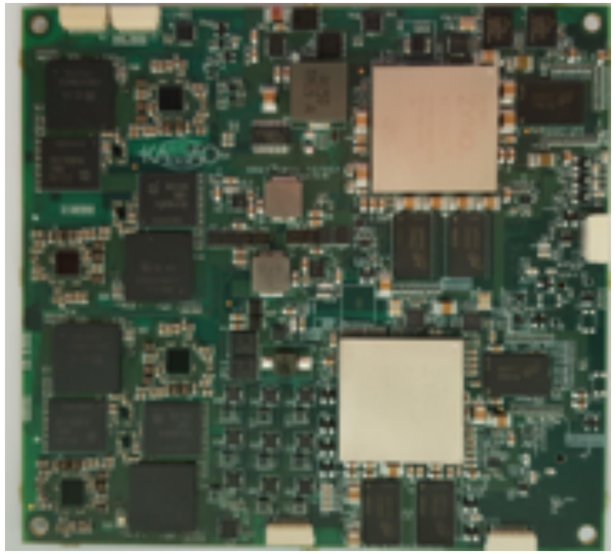
NUMASCALE Architecture Overview

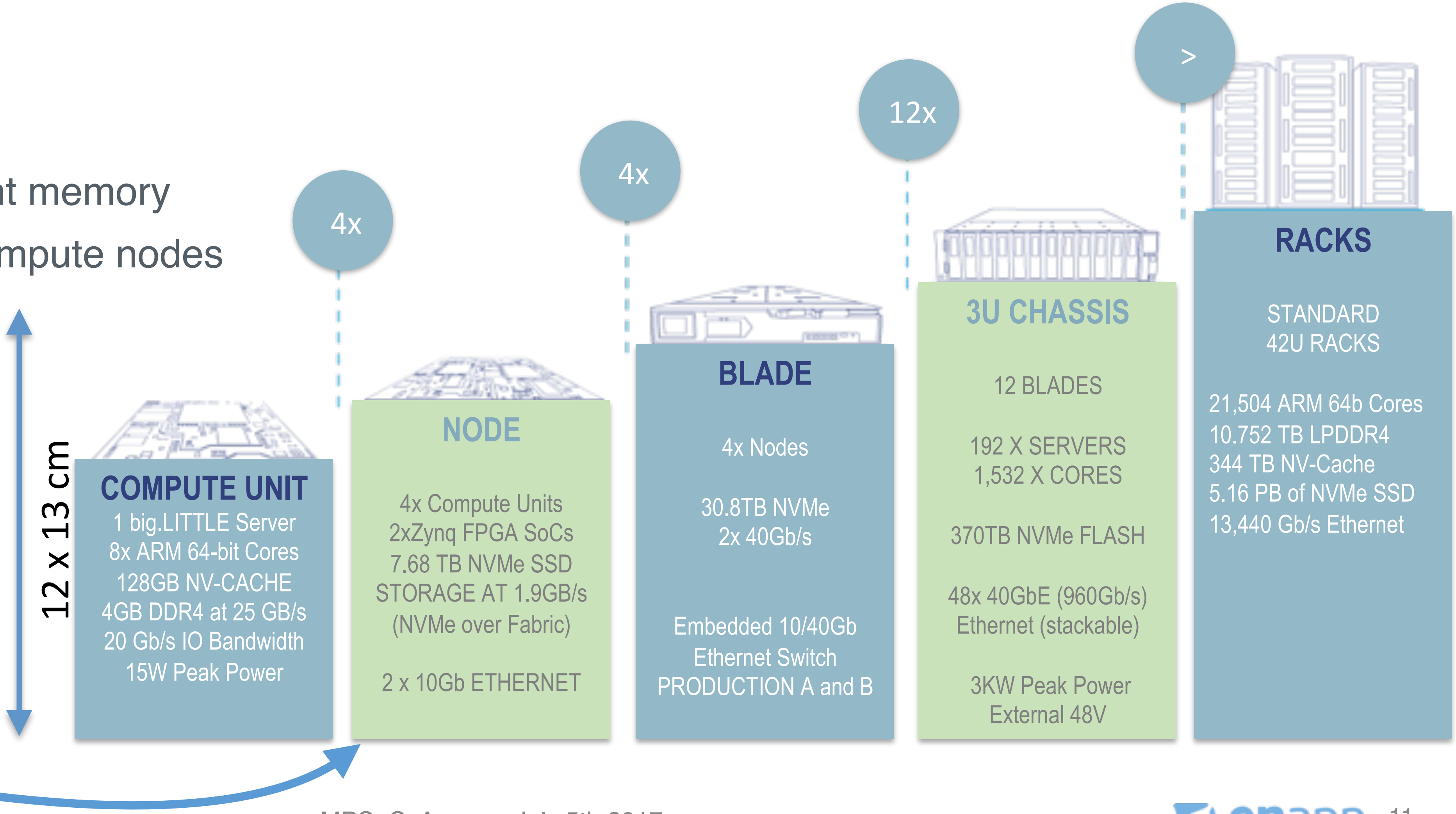


- Multi-node clustering vs Numachip
- Aggregate resources on the HW level
- Cache-coherent multi-node systems
- Single OS to handle all clustered resources
- Share everything

KALEAO Integrated PCB (Compute Node)

- Hardware accelerated I/O
- Low-power
- Share-nothing
- UNIMEM coherent memory access across compute nodes





COMPUTE UNIT

1 big.LITTLE Server
8x ARM 64-bit Cores
128GB NV-CACHE
4GB DDR4 at 25 GB/s
20 Gb/s IO Bandwidth
15W Peak Power

NODE

4x Compute Units
2xZynq FPGA SoCs
7.68 TB NVMe SSD
STORAGE AT 1.9GB/s
(NVMe over Fabric)

2 x 10Gb ETHERNET

BLADE

4x Nodes

30.8TB NVMe
2x 40Gb/s

Embedded 10/40Gb
Ethernet Switch
PRODUCTION A and B

3U CHASSIS

12 BLADES

192 X SERVERS
1,532 X CORES

370TB NVMe FLASH

48x 40GbE (960Gb/s)
Ethernet (stackable)

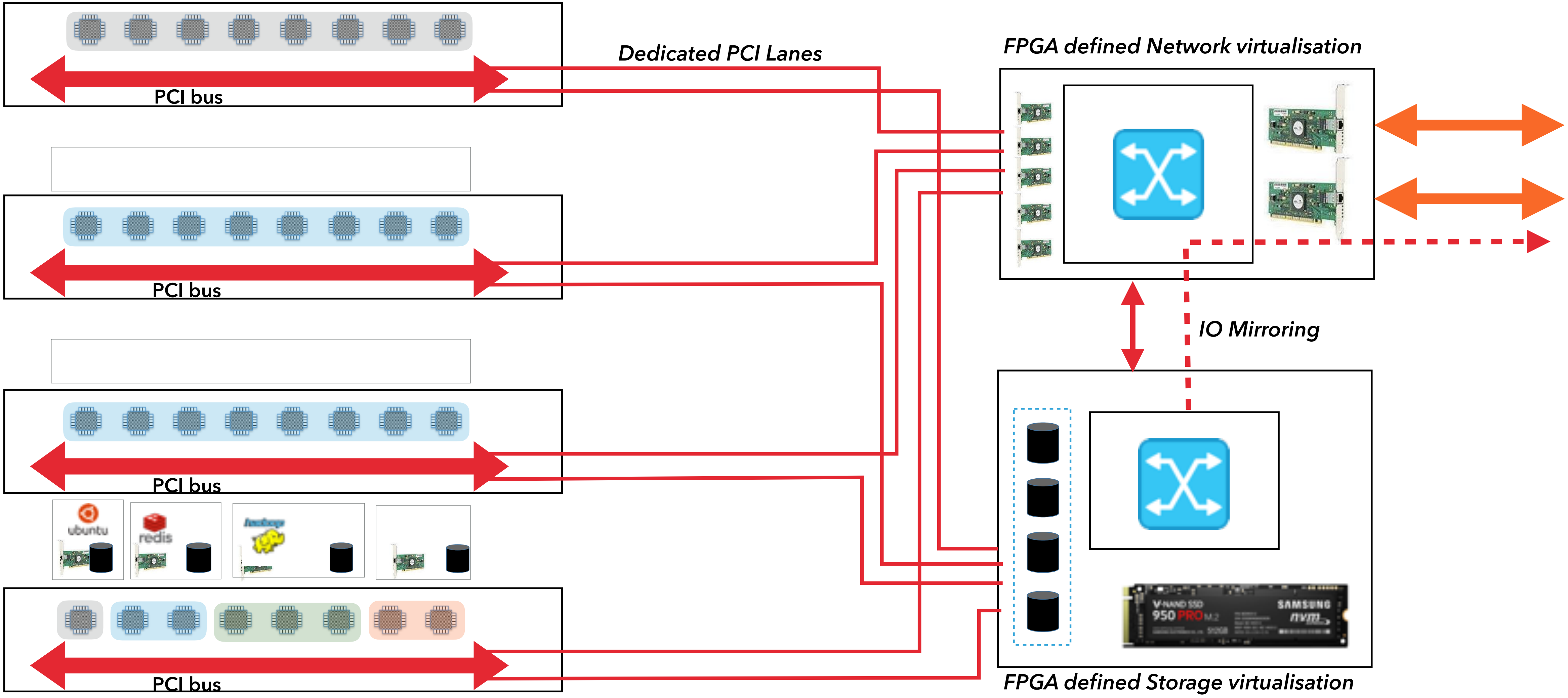
3KW Peak Power
External 48V

RACKS

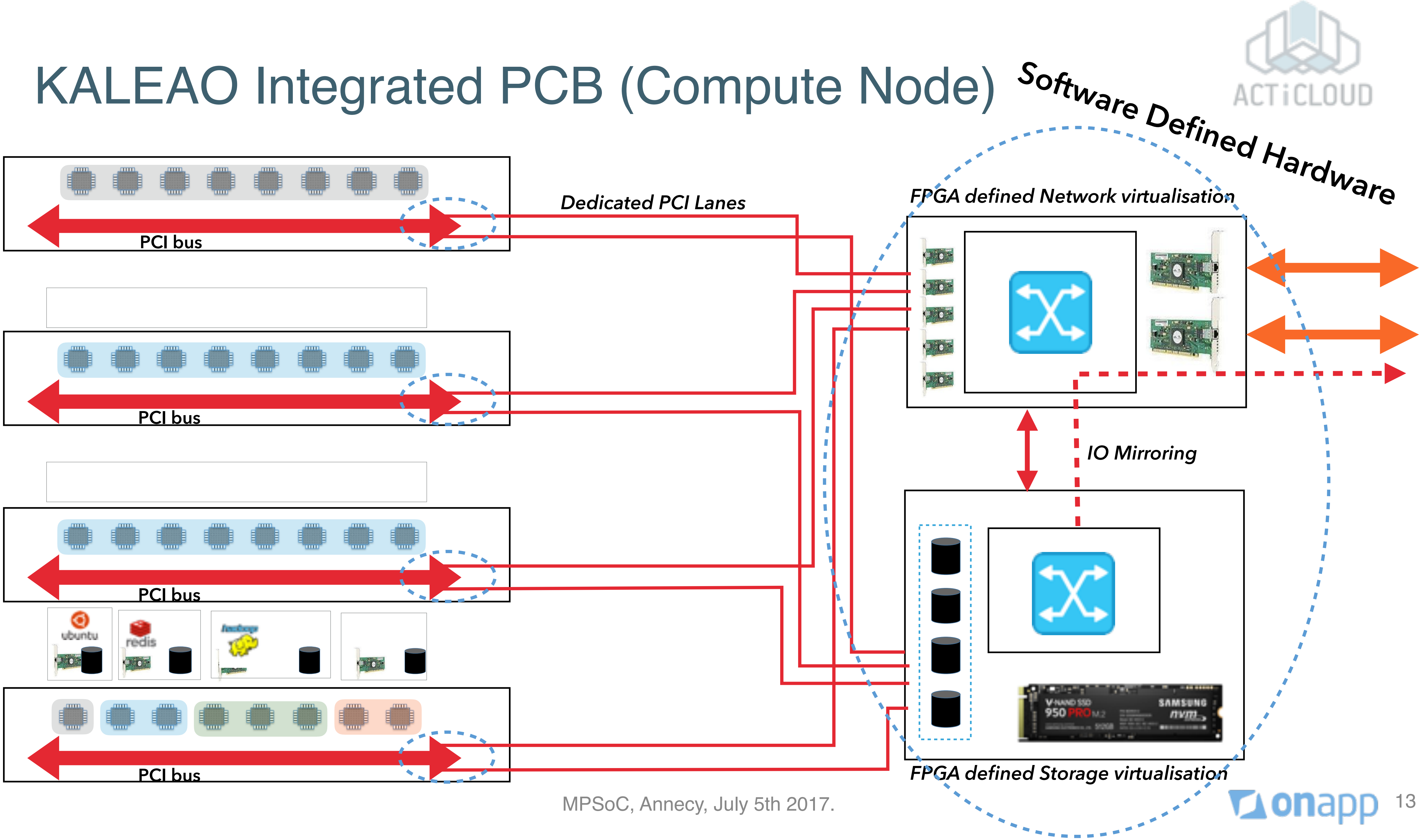
STANDARD
42U RACKS

21,504 ARM 64b Cores
10.752 TB LPDDR4
344 TB NV-Cache
5.16 PB of NVMe SSD
13,440 Gb/s Ethernet

KALEAO Integrated PCB (Compute Node)



KALEAO Integrated PCB (Compute Node)



Multi-tenancy in the DC

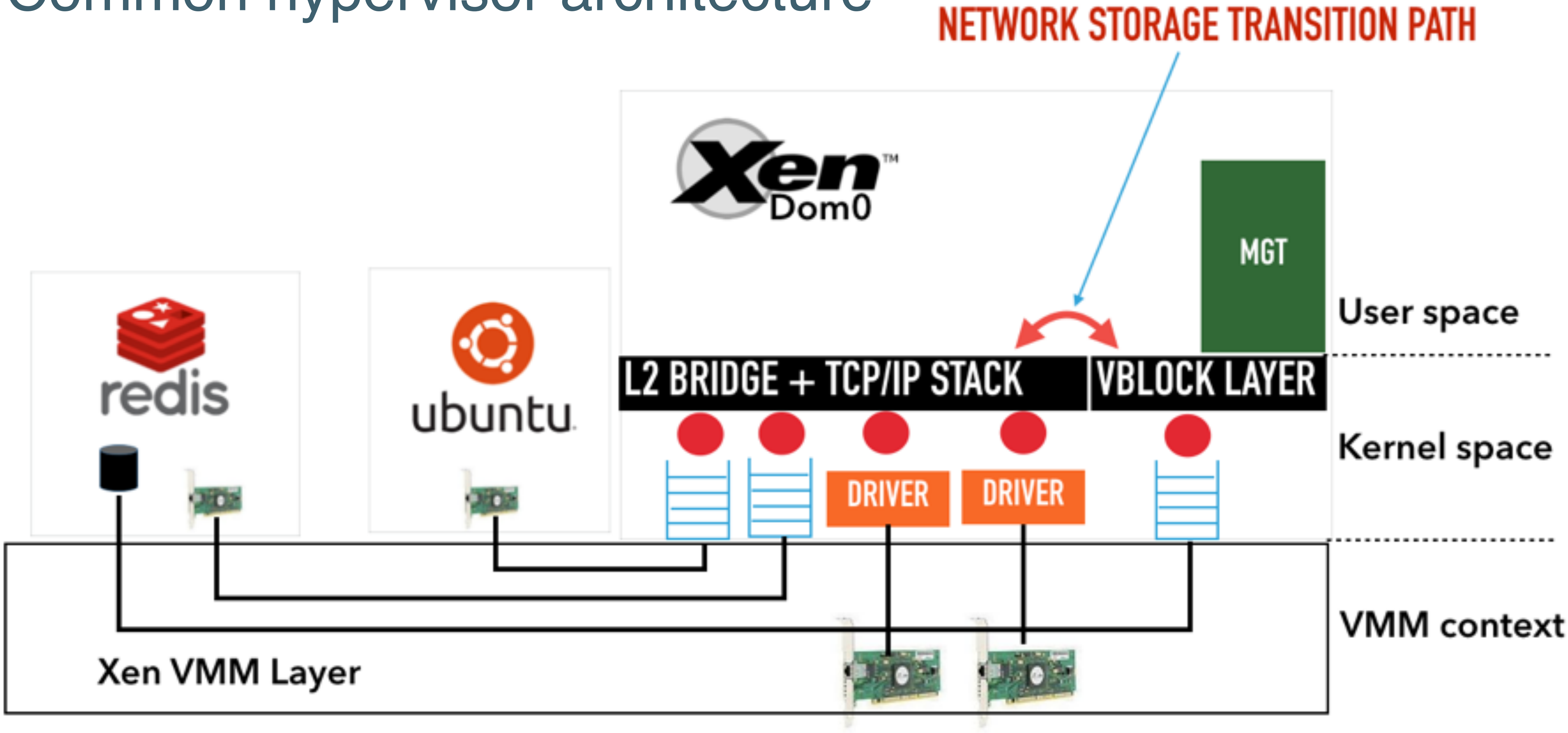
- Multi-tenant server operation is **ubiquitous** in the modern DC
  - efficient utilisation of hardware resources
  - high availability/Disaster Recovery for virtual server workloads requires redundant infrastructure and motion of workloads
- Traditional hypervisor architecture is optimised towards large Intel NUMA systems
  - large footprint control domain with full TCP/IP stack management interfaces
  - all virtual IO queues are multiplexed through the hostOS
  - 1-2GB memory footprint + 2 or more physical cores reserved just for management domain

Designing a rackscale low power SoC Hypervisor

System software architecture

- Clustered Hypervisor technology, centrally managed nodes with no control domain on each
  - Scale up to many thousands of managed nodes from a single controller
  - Very lightweight raw ethernet-based management interface
- Designed to integrate seamlessly with FPGA co-processor(s) for IO management
  - ***Software Defined hardware acceleration***
- Based on Xen, with a complete re-architecture of VMM IO subsystem and the management/control interface
  - Achieves native hardware IO performance for VMs
- Super-fine grained resource management per core/socket/controller/memory address

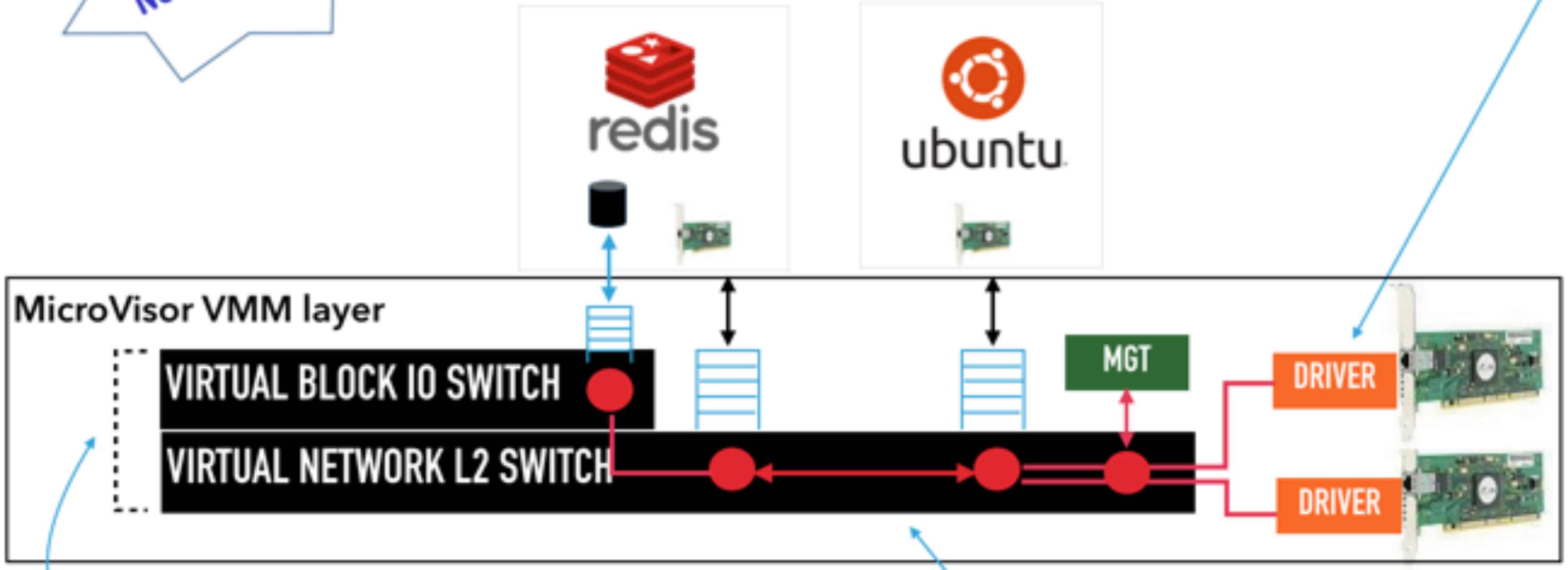
Common hypervisor architecture



MicroVisor integrated architecture

NO DOMAIN 0

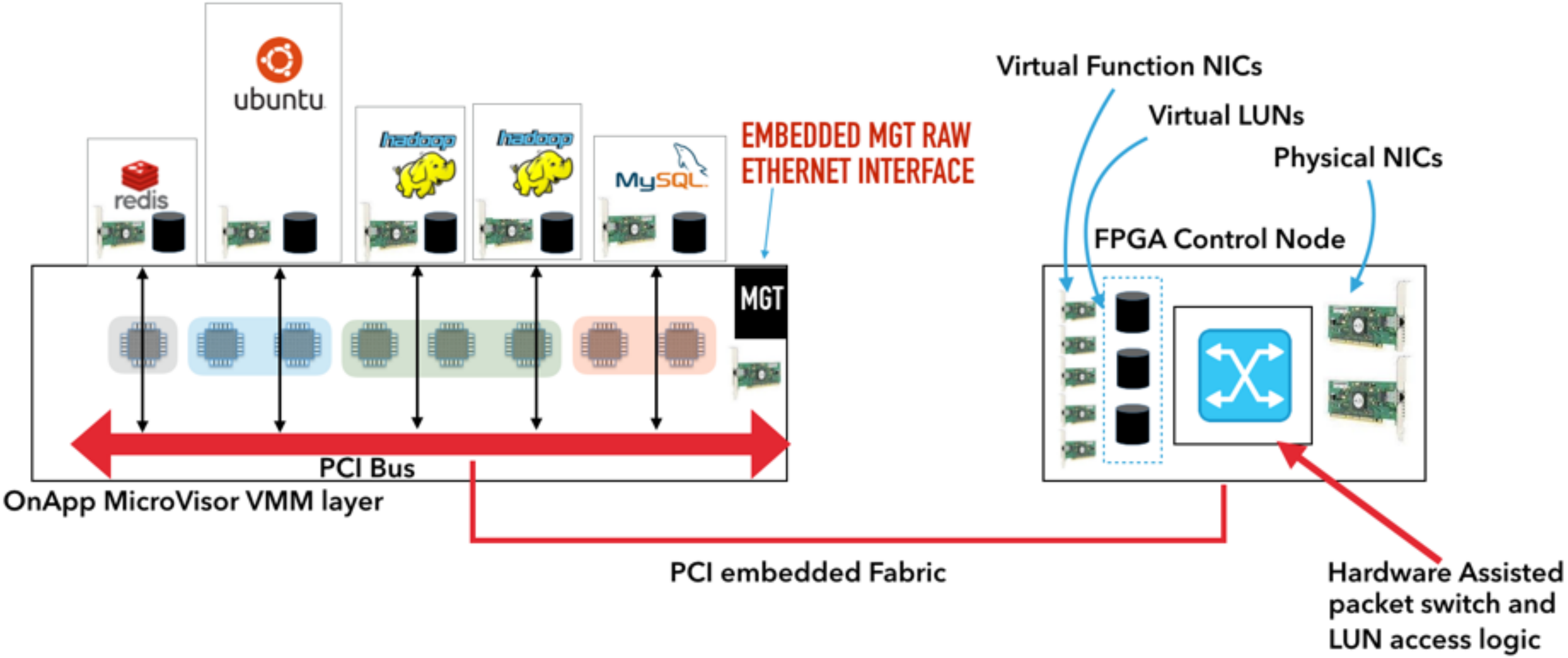
EMBEDDED NIC DRIVER INTEGRATED WITH VIRTUAL PACKET SWITCH



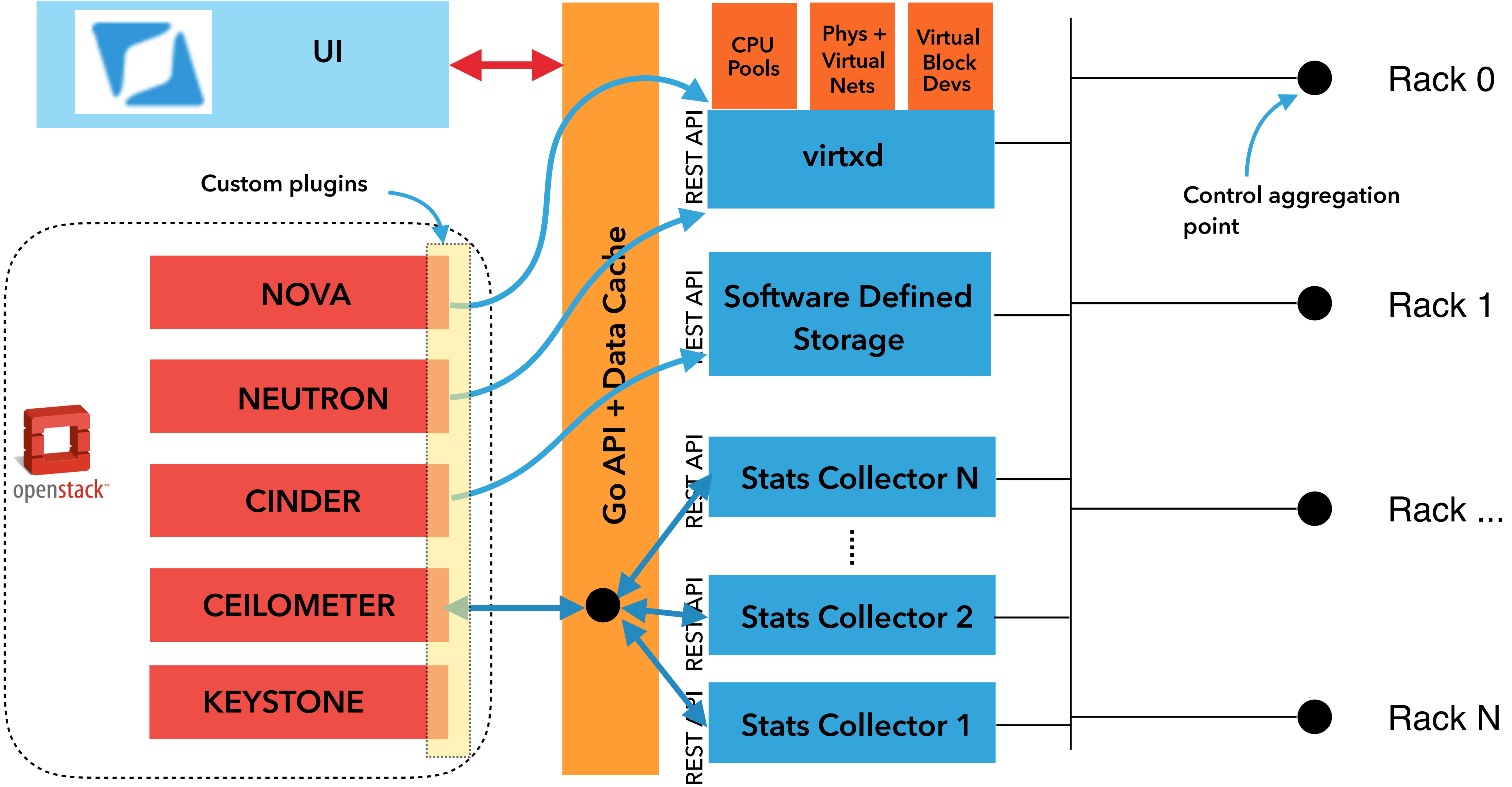
INTEGRATED BLOCK AND L2 PACKET SWITCH

VIRTUAL IO QUEUES TIGHTLY INTEGRATED TO HARDWARE DMA

FPGA Acceleration Integration - Software Defined Hardware



MicroVisor Management



Software Defined Hardware - accelerating distributed block storage

MPSoC, Annecy, July 5th 2017.

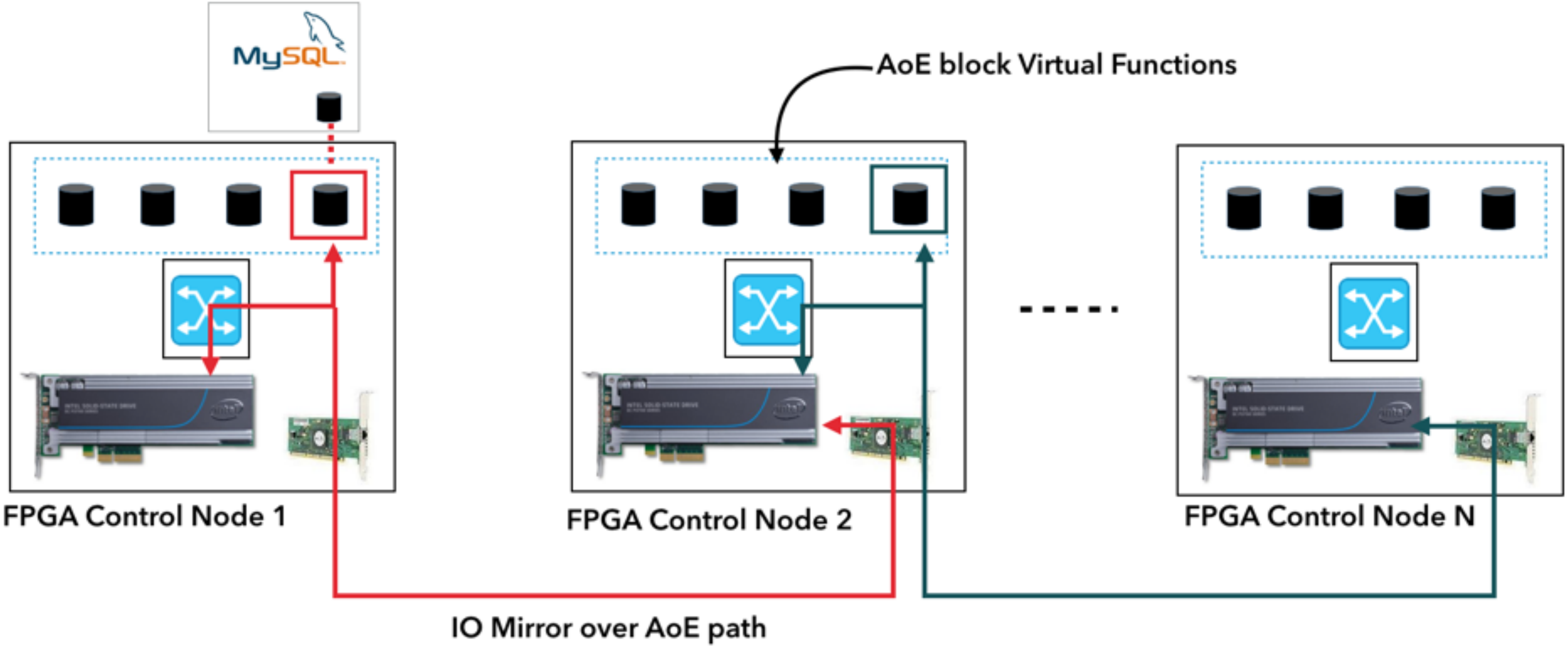
OnApp SDS technology today

- Hyper-converged storage solution, built for the OnApp cloud platform
- Each Hypervisor advertises and enables remote access to direct attached storage drives
  - Block path frontend mirrors data across both local and remote paths
  - Failures are tolerated at frontend and resynched in the background across controllers independently
- TCP or ATA over Ethernet protocol used for fast bock access between nodes
- Transparent data relocation/content balancing provided whilst VMs stay online
- Scales across 100s of physical nodes (1000s of drives)
- Thin provisioning, fast snapshot and clone, wide area data replication are standard

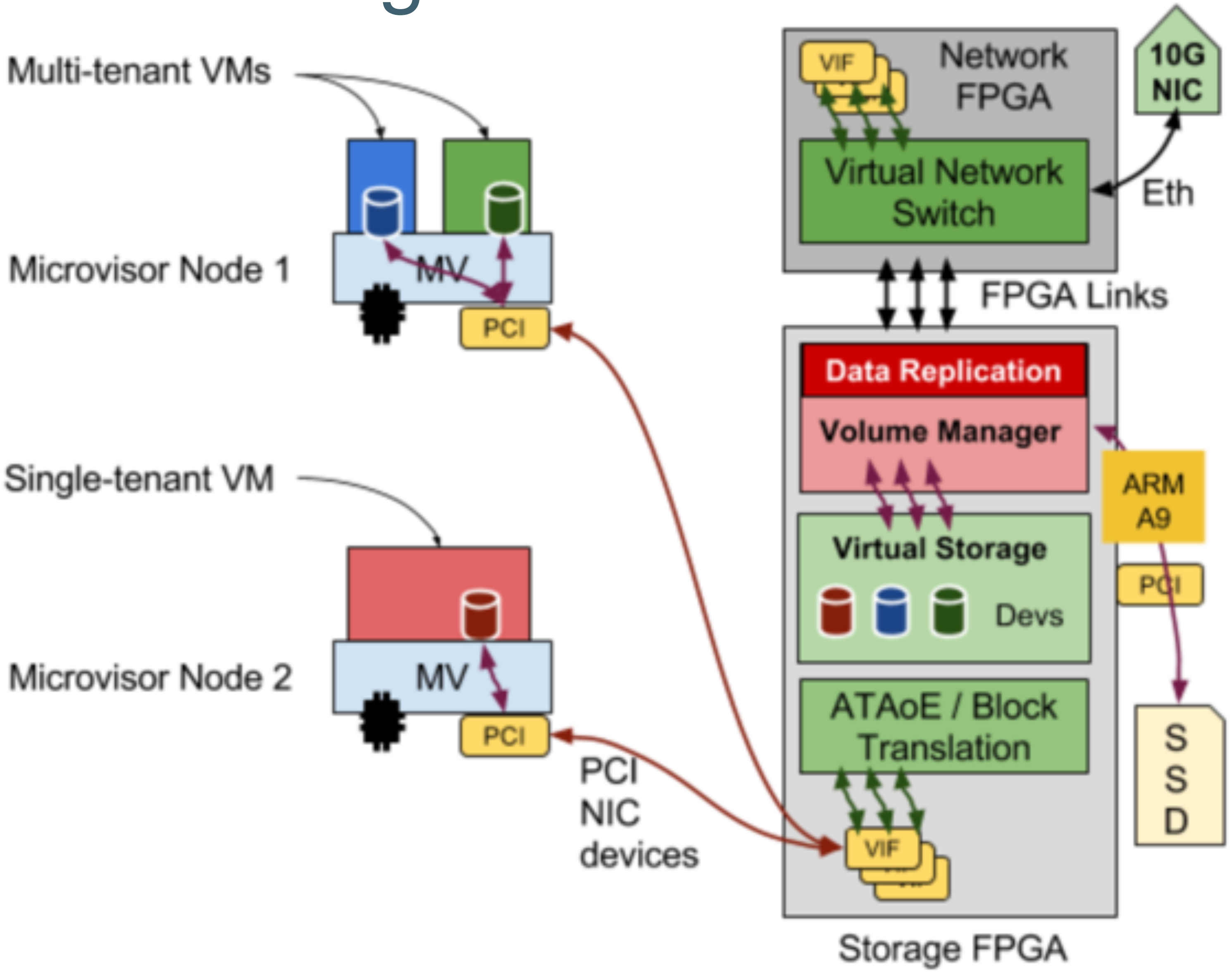
Offloading OnApp SDS into an FPGA

- Each FPGA unit directly manages physical NVMe storage
- Lightweight linux management host runs on the embedded ARM cores of the Zynq FPGA processor
    - control stack for the hardware node
    - manage the allocation bitmap for virtual LUNs hosted on the local NVMe storage
    - signal the virtual to physical block map tables to the FPGA
- the FPGA programmable logic handles ATAoE frames directly and maps to/from the NVMe storage
- AoE client signals to the FPGA device extra attributes:
    - Data mirror list for IO writes
    - Data copy command + destination address for resynch of data

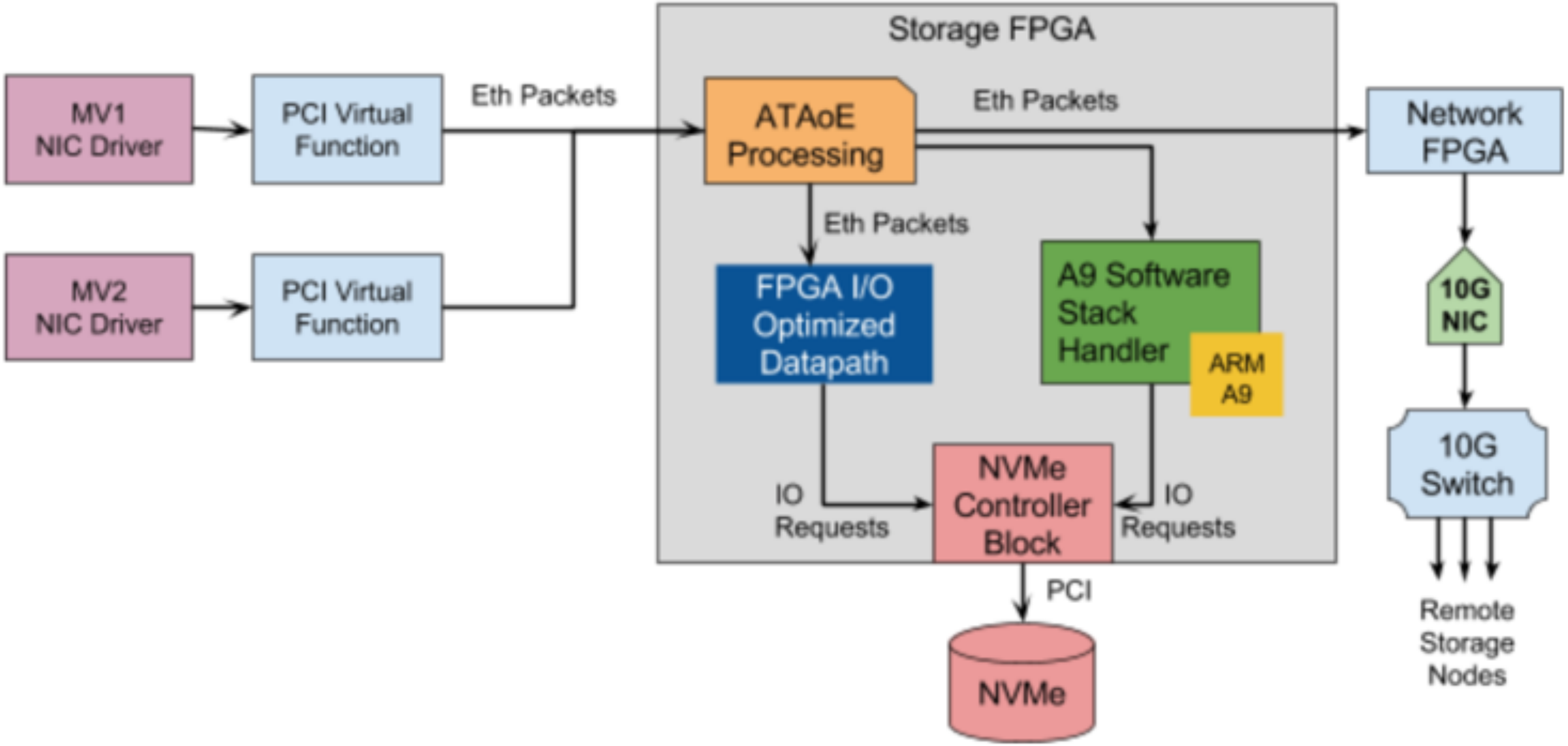
Integrated NVMe over Fabric



Storage FPGA Logical Elements



Logical Packet Processing Flow



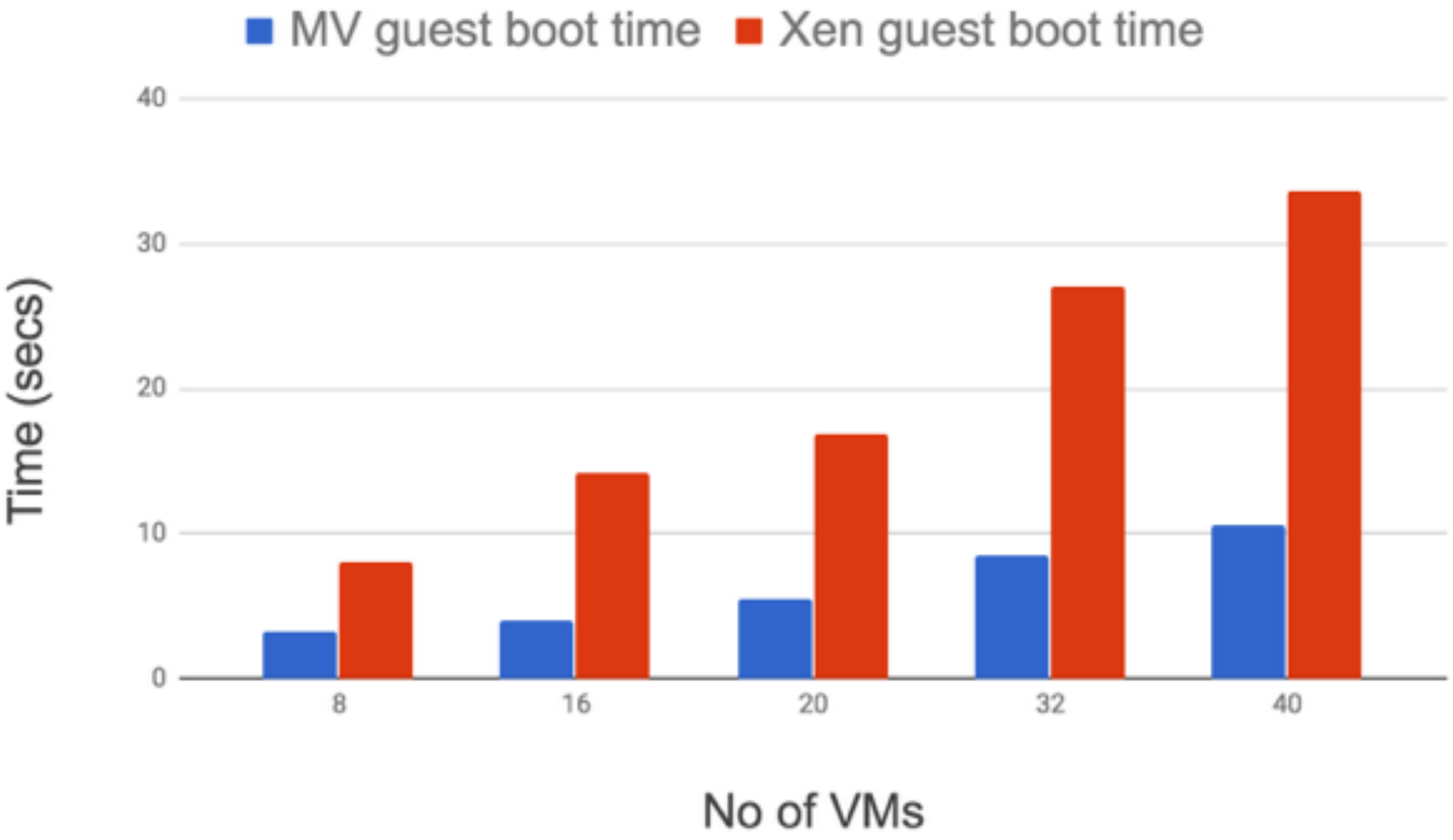
Hardware/Software Co-design

- Software client is responsible for runtime signalling
  - extended packet header provides replication MAC address lists
  - packet type indicates READ, WRITE or COPY operation
  - path failure detection handled in software on client side
- A9 control system is responsible for data path setup and volume management
  - slow path for block requests that are not provisioned
  - thin provisioned V2P table updated dynamically
  - executes all the SDS content distribution algorithms
- FPGA PL responsible for fast data path handling
  - process AoE block requests directly to the NVMe storage
  - mirror packets to remote nodes based on packet header lists
  - copy data and forward to remote nodes for fast re-synchronisation of data

Performance Benefits

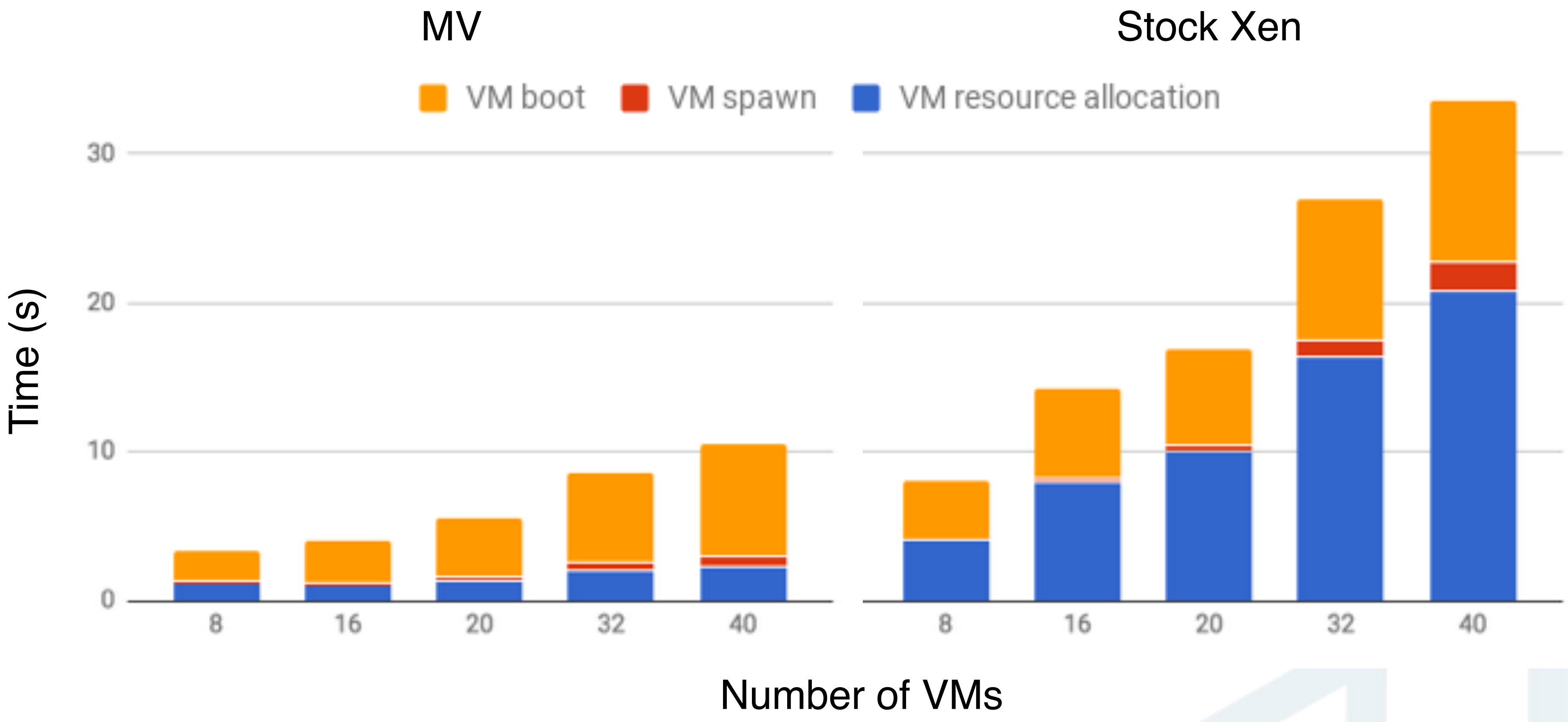
MicroVisor guest Boot time (vs Stock Xen)

- spawn guests in parallel
- start timer at spawn
- stop timer at first ping from the guest (triggered from the last service in the boot chain)

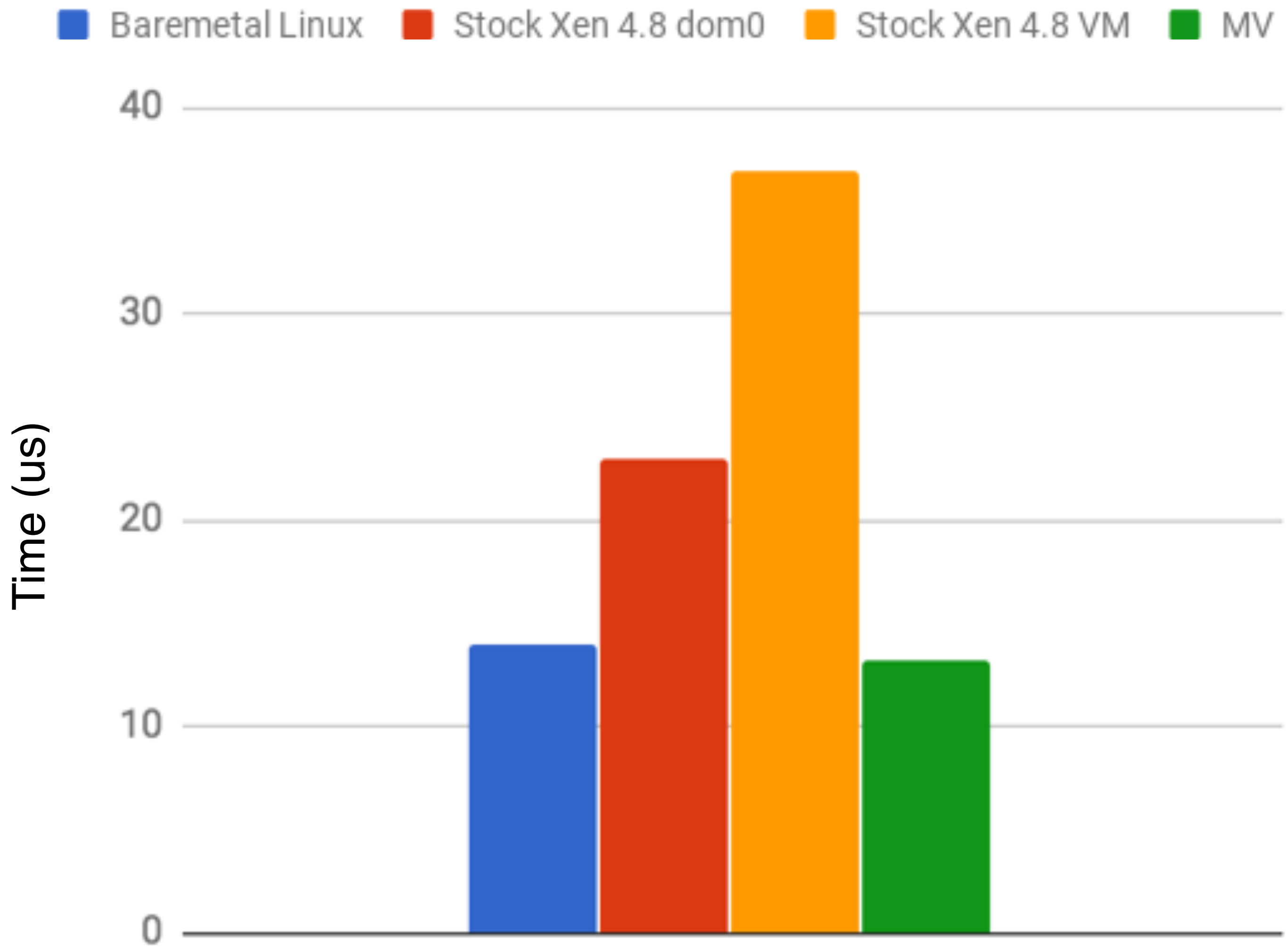


MPSoC, Annecy, July 5th 2017.

VM boot time breakdown



Latency



- 1-way latency
- No TCP/UDP protocols involved — custom raw ethernet latency tool
- Off-the-shelf Intel 10GbE

Conclusions

- Many core, integrated low power SoC designs are coming to cloud scale DCs
- FPGA acceleration technology features prominently in the roadmap for the DC
  - leveraging hardware acceleration is critical in achieving native hardware performance in upcoming IO interface advances

onapp

- OnApp has designed and built a clustered Hypervisor that is designed to support thousands of integrated low power SoC processing nodes with minimal control and management overhead on each node
  - Management system overhead per coherent node is an order of magnitude smaller than a traditional hypervisor system
  - IO architecture is optimised to move IO much more efficiently to centralised hardware processing units
- OnApp is leveraging FPGA acceleration technology to build a hybrid Software Defined Hardware Accelerated Distributed Storage technology

Thanks!

More info:

julian@onapp.com

https://onapp.com

https://acticloud.eu

@ACTiCLOUD

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement no 732366 (ACTiCLOUD)