(MRC)^2

Modular Research-based Composably trustworthy Mission-oriented Resilient Clouds

or…

Scaling to a million switchlets?

Peter G. Neumann
SRI International

Simon W. Moore
Robert Watson
University of Cambridge

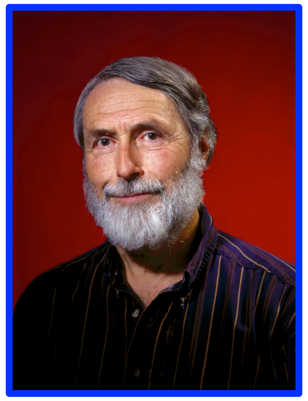
MRC PI Meeting
San Diego, CA
30 October 2012

This effort is sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL), under contract FA8750-11-C-0249. This material is approved for public release, distribution unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

SRI International®

UNIVERSITY OF CAMBRIDGE

The $(MRC)^2$ team



Dr Peter G. Neumann

Dr Robert N.M. Watson

Dr Simon W. Moore

Dr Nirav Dave

Mr Brooks Davis

Dr Hassen Saidi

Dr Patrick Lincoln

Mr Phillip Porras

Mr Stacey Son

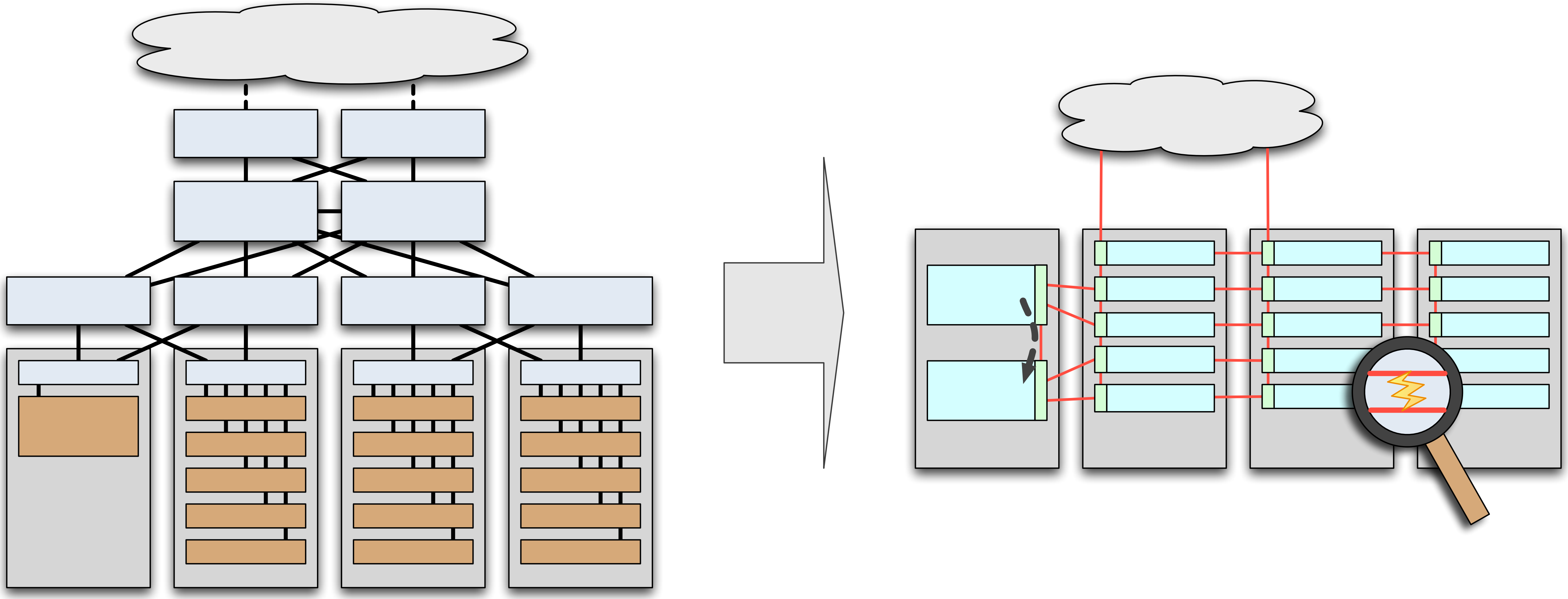
🟦 SRI International

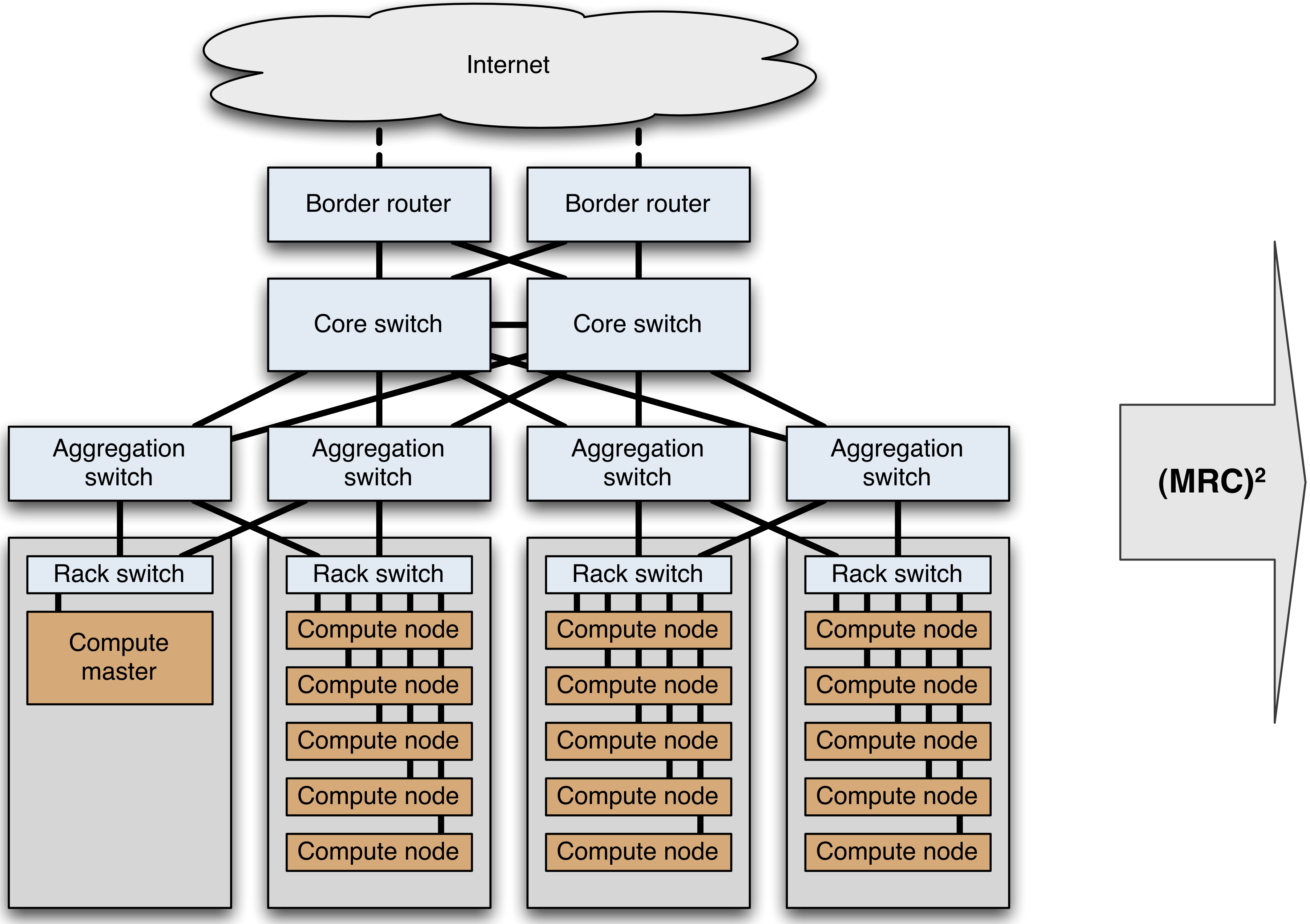
🟥 University of Cambridge

MRC2 project members unable to attend the PI meeting:

Jonathan Anderson, David Chisnall, Matthew P. Grosvenor, Khilan Gudka, Asif Khan, Myron King, Anil Madhavapeddy, Andrew Moore Alan Mujumdar, Steven J. Murdoch, Robert Norton, Muhammad Shahbaz, Richard Uhler, Jonathan Woodruff, Vinod Yegneswaran, Dongting Yu

(MRC)² data centre





Internet

Border router

Border router

Core switch

Core switch

Aggregation switch

Aggregation switch

Aggregation switch

Aggregation switch

Rack switch

Compute master

Rack switch

Compute node

Compute node

Compute node

Compute node

Compute node

Rack switch

Compute node

Compute node

Compute node

Compute node

Compute node

Rack switch

Compute node

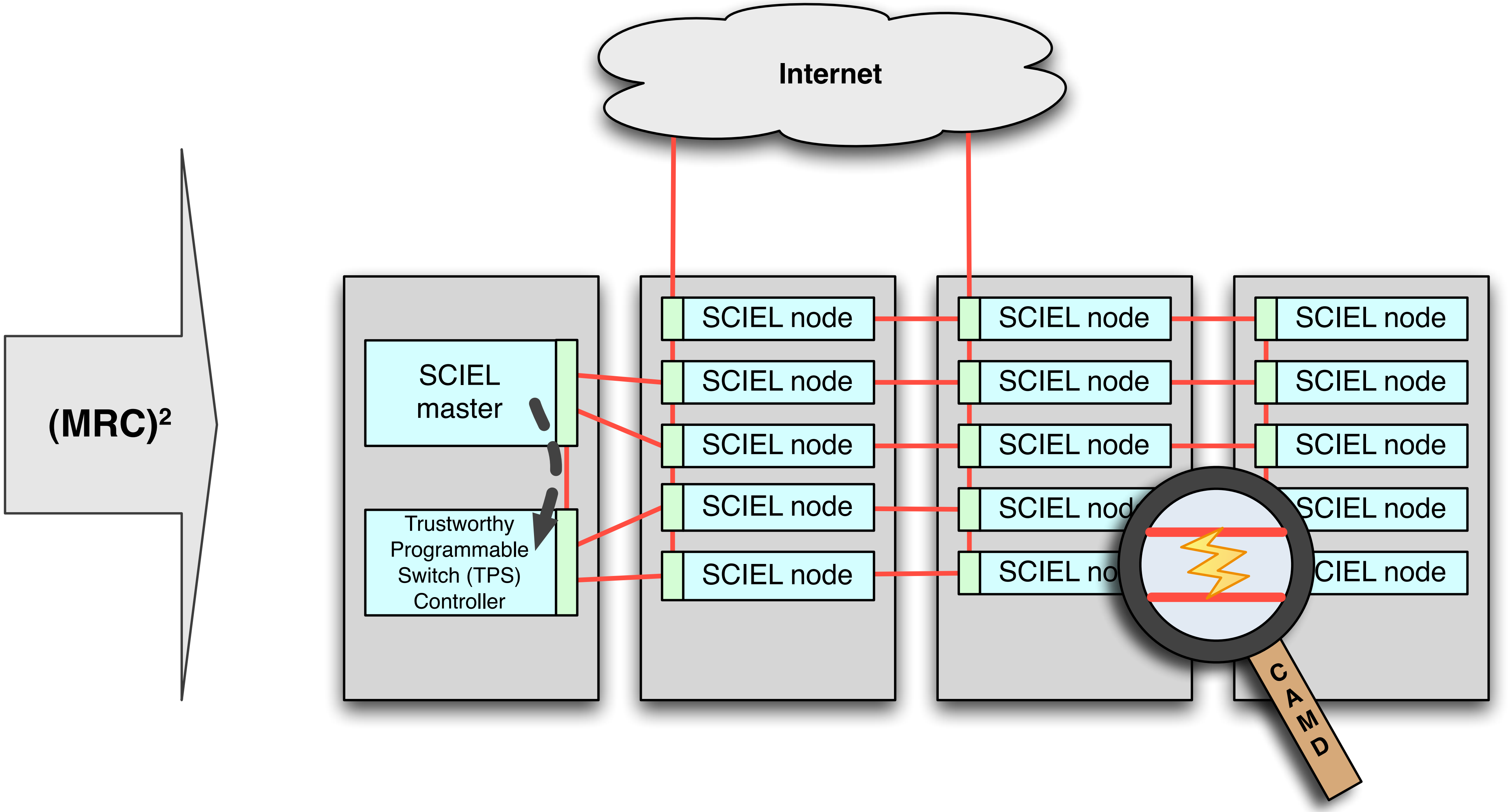
Compute node

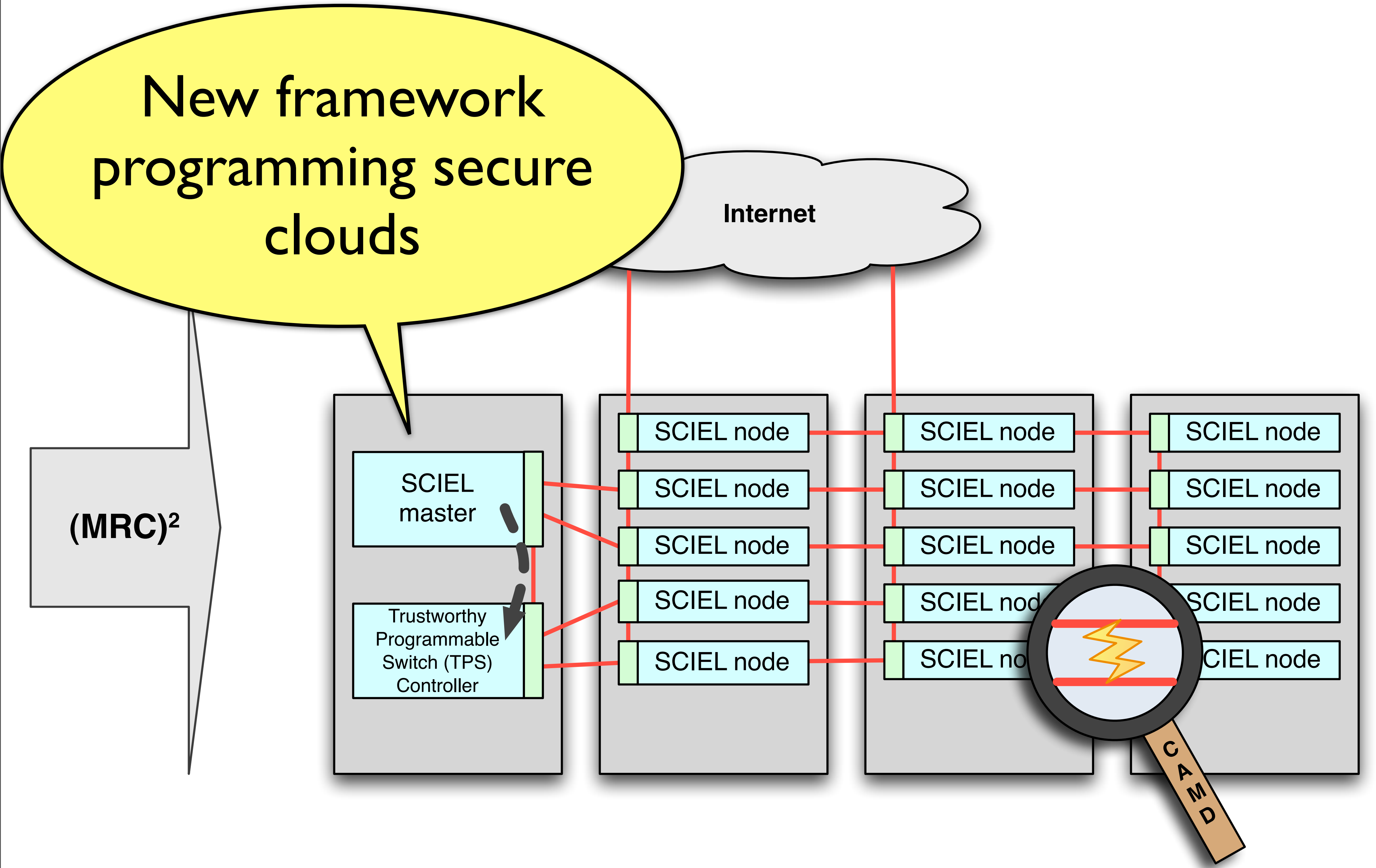
Compute node

Compute node

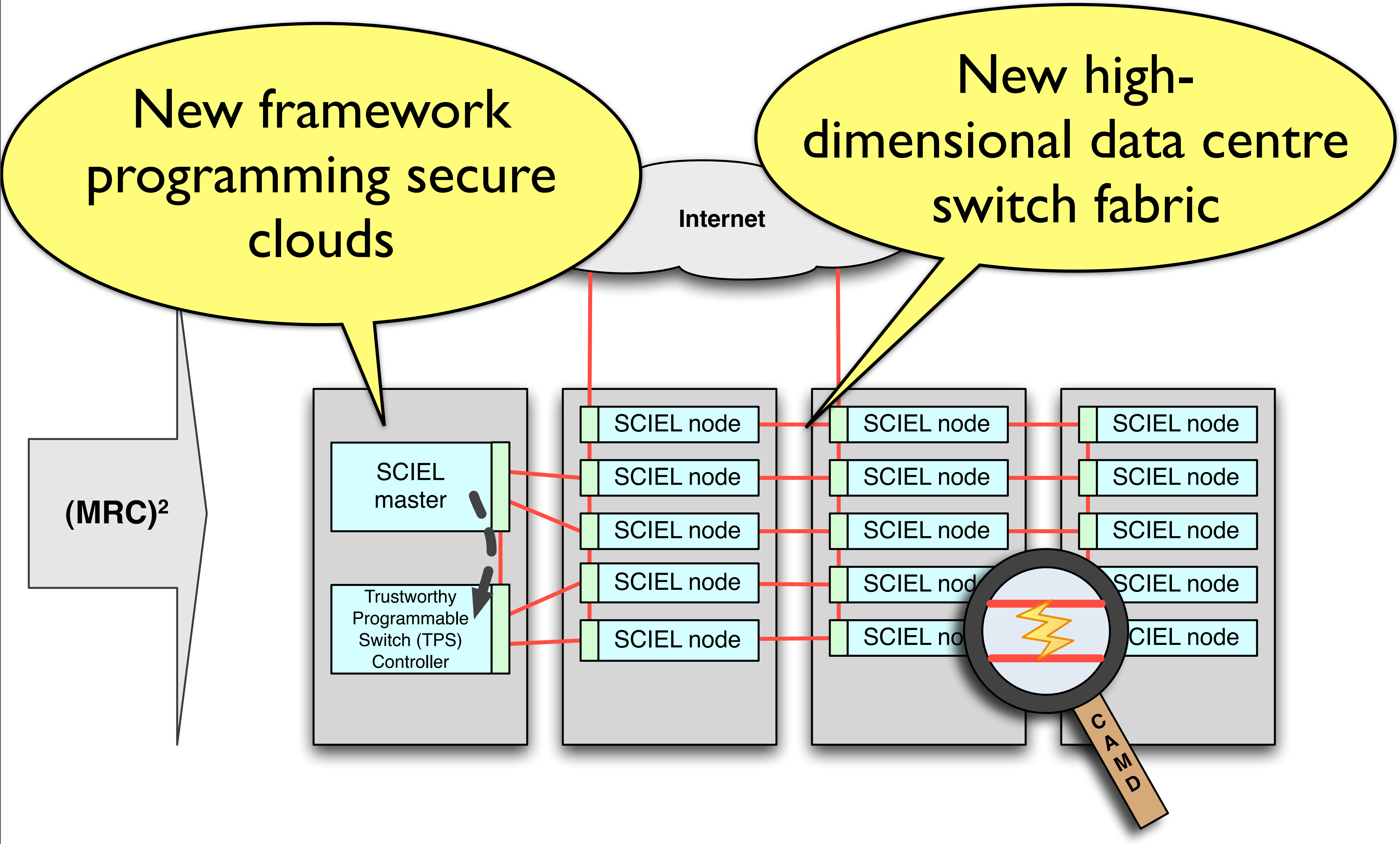
Compute node

(MRC)²


(MRC)²
Internet
SCIEL master
Trustworthy Programmable Switch (TPS) Controller
SCIEL node
SCIEL node
SCIEL node
SCIEL node
SCIEL node
SCIEL node
SCIEL node
SCIEL node
SCIEL nod
SCIEL no
SCIEL node
SCIEL node
SCIEL node
SCIEL node
CIEL node
CAMD

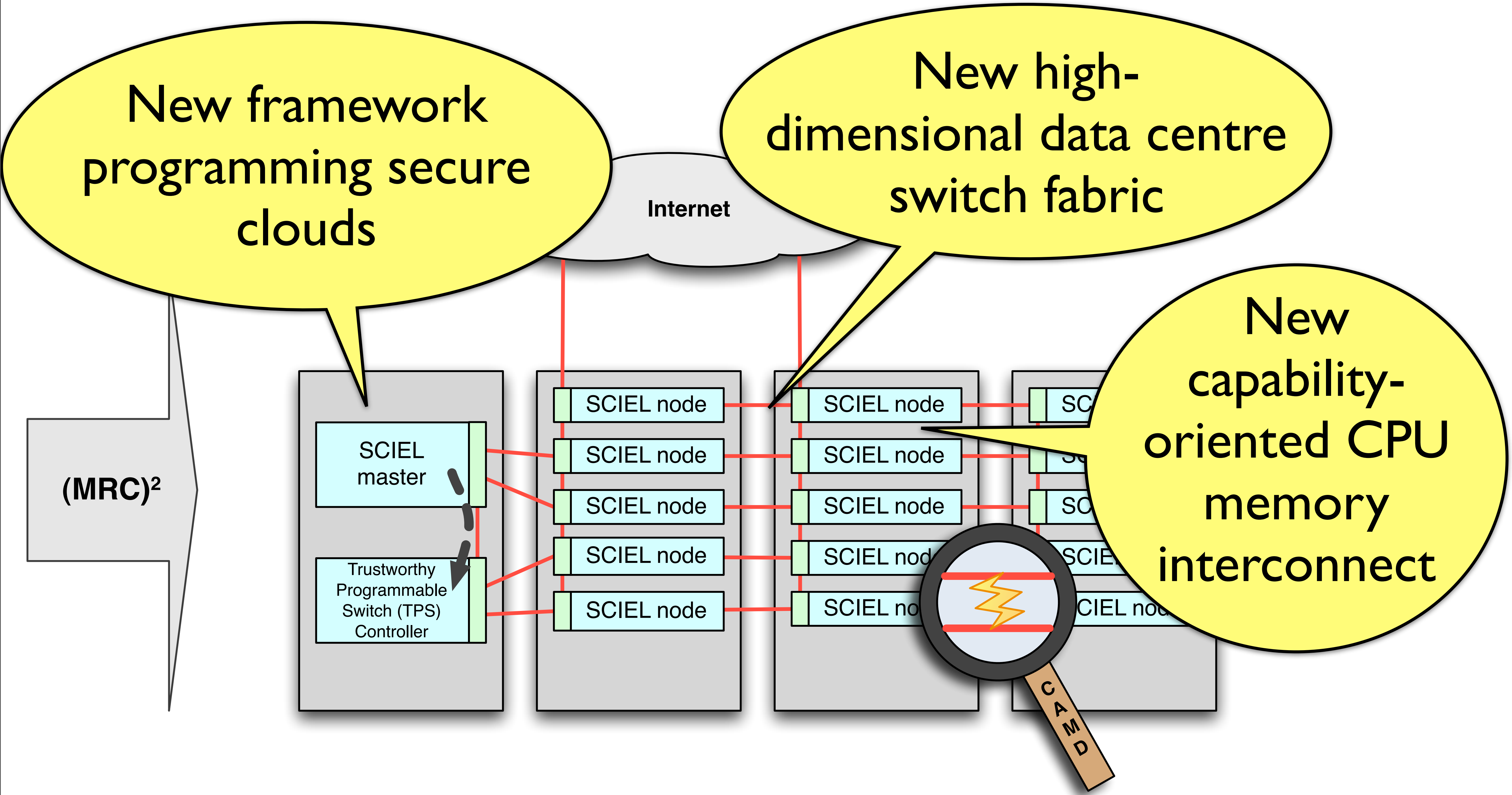


New framework programming secure clouds



New framework programming secure clouds

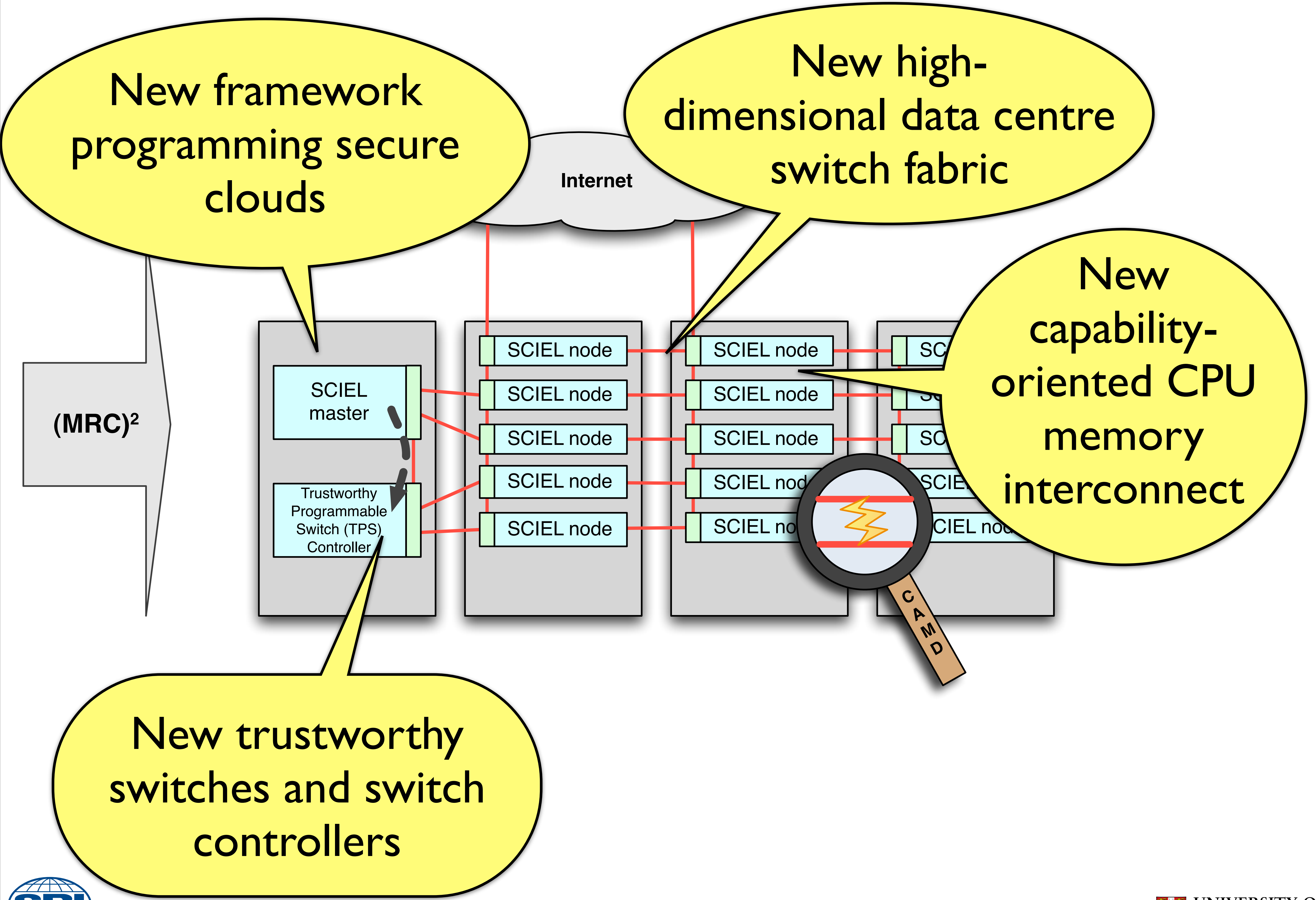
New high-dimensional data centre switch fabric



New framework programming secure clouds

New high-dimensional data centre switch fabric

New capability-oriented CPU memory interconnect



New framework programming secure clouds

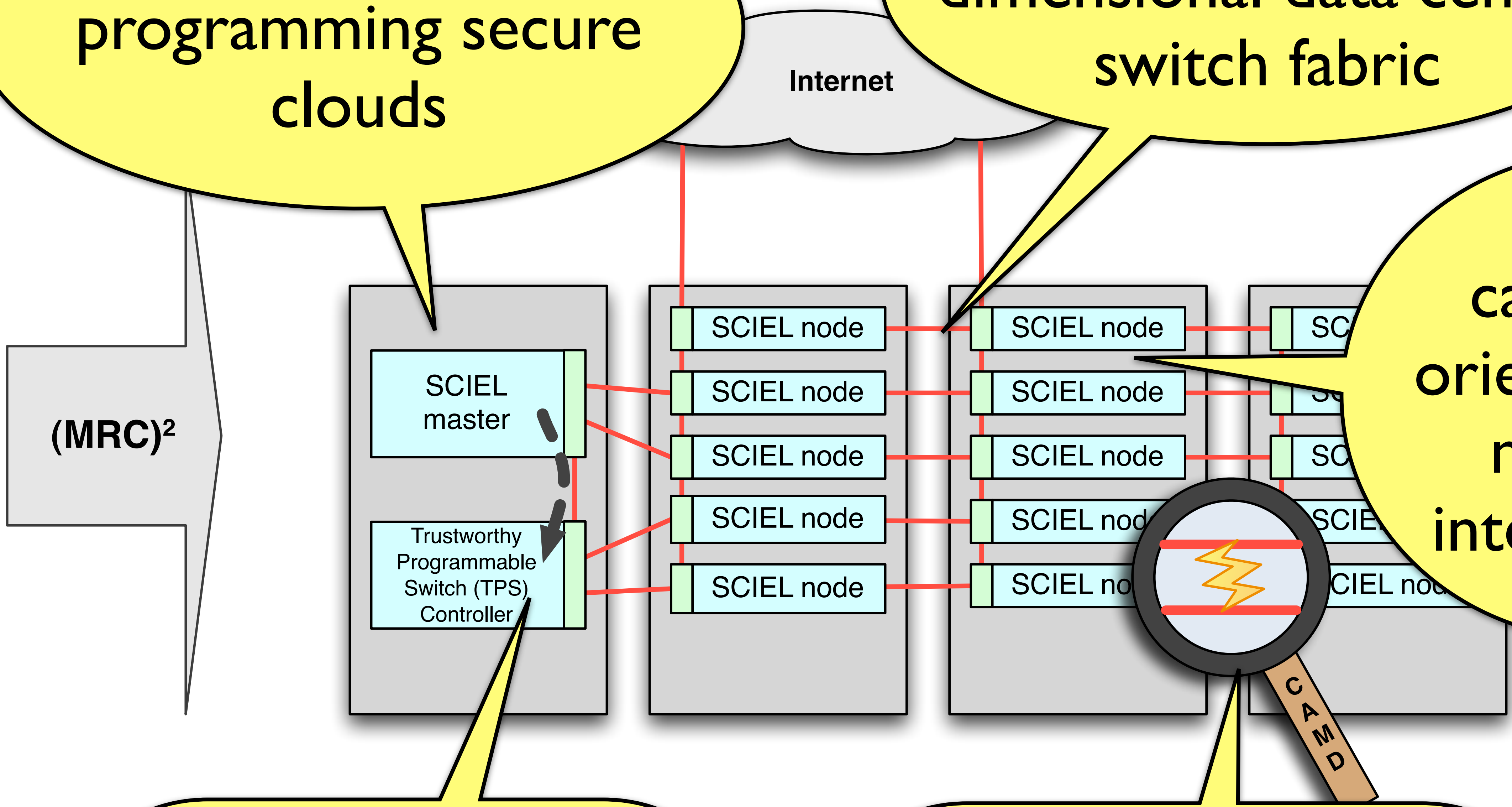
New high-dimensional data centre switch fabric

New capability-oriented CPU memory interconnect

New trustworthy switches and switch controllers

SRI International

UNIVERSITY OF CAMBRIDGE



New framework programming secure clouds

New high-dimensional data centre switch fabric

New capability-oriented CPU memory interconnect

New trustworthy switches and switch controllers

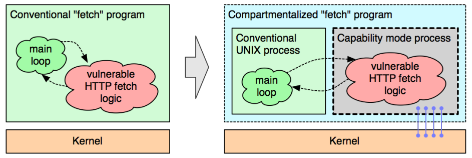
New distributed analysis, analytics, and misuse detection and response

Cross-cutting themes

- Data centre switching
- Distributed resilience throughout
- Aligning algorithm and network topology
- Energy-efficiency/security/resilience/scalability tradeoffs
- Multi-scale computing techniques
- Capability system security models
- Formal grounding

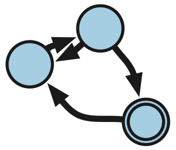
Research inputs to MRC2

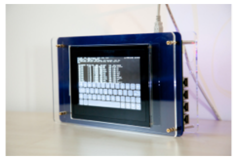
Capsicum

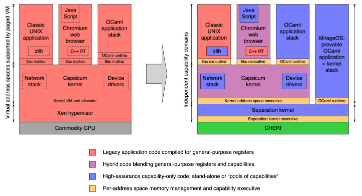

Conventional "fetch" program
main loop
vulnerable HTTP fetch logic
Kernel
Compartmentalized "fetch" program
Conventional UNIX process
Capability mode process
main loop
vulnerable HTTP fetch logic
Kernel

Hybrid capability models
Application compartmentalization

CTSRD






Virtual address spaces supported by paged VM
Classic UNIX application
zlib
libc malloc
Java Script
Chromium web browser
C++ RT
libc malloc
OCaml application stack
OCaml runtime
libc malloc
Network stack
Capsicum kernel
Device drivers
Kernel VM and allocator
Xen hypervisor
Commodity CPU
Independent capability domains
Classic UNIX application
zlib
libc executive
Java Script
Chromium web browser
C++ RT
libc executive
OCaml application stack
OCaml runtime
MirageOS: provable OCaml application + kernel stack
Network stack
Capsicum kernel
Device drivers
Kernel address space executive
OCaml runtime
Separation kernel
Separation kernel executive
CHERI
Legacy application code compiled for general-purpose registers
Hybrid code blending general-purpose registers and capabilities
High-assurance capability-only code; stand-alone or "pools of capabilities"
Per-address space memory management and capability executive

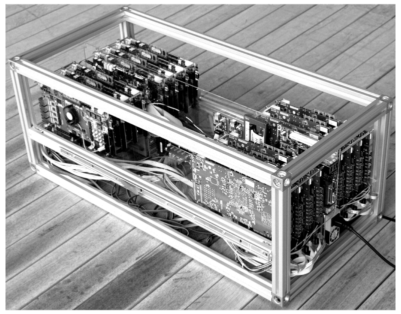
BERI hardware-software research platform
CHERI hybrid-capability protection model
Bluespec formal methodology
Temporal assertion techniques

Cyber Threat Analytics

FortNox SDN TCB
SDN security applications

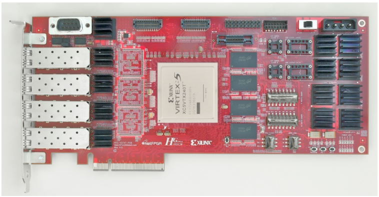
(MRC)^2

BIMPA



Multi-FPGA hardware
prototyping platform

NetFPGA



Switch, OpenFlow, NIC
hardware research platform

Mirage a cloud operating system

A secure, high-performance
exokernel written in OCaml

SRI International

UNIVERSITY OF CAMBRIDGE

(MRC)^2 research topics

Chimera	Rack-scale capability-oriented memory interconnect
RDSF	Higher dimensional data centre switching
TPSC	Trustworthy, distributed Software-Defined Networking (SDN) controllers
CAMD	Cloud analysis and misuse detection
SCIEL	A programming framework for secure resilient clouds

UNIVERSITY OF CAMBRIDGE

Chimera

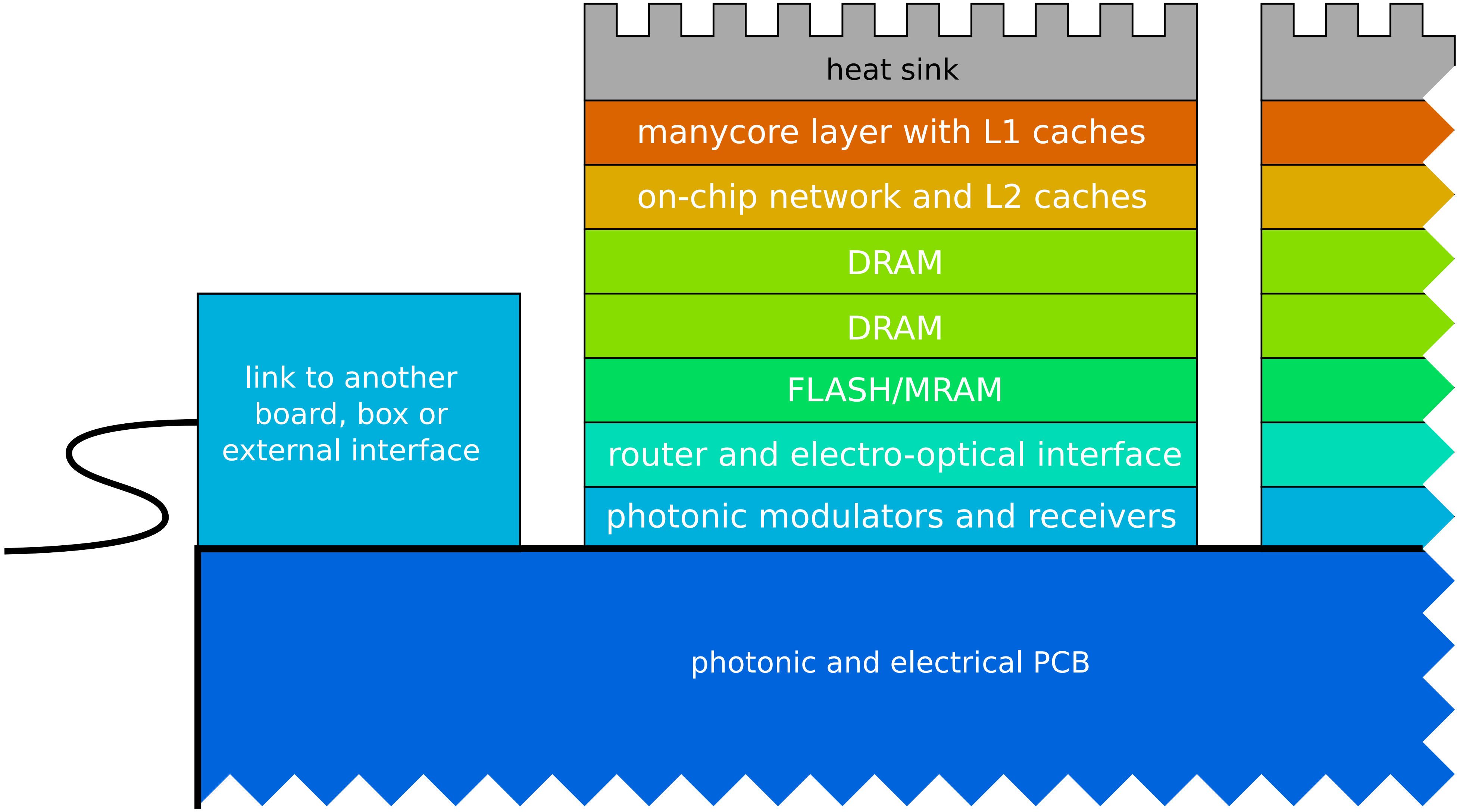
Rack-scale capability-oriented memory interconnect

- Build on single-core CHERI processor from CTSRD project in CRASH program
- Investigate capabilities to manage information flow
  - More scalable protection model
  - Exploit additional memory semantics visible to CPU
  - Explore consistency effects on capability model

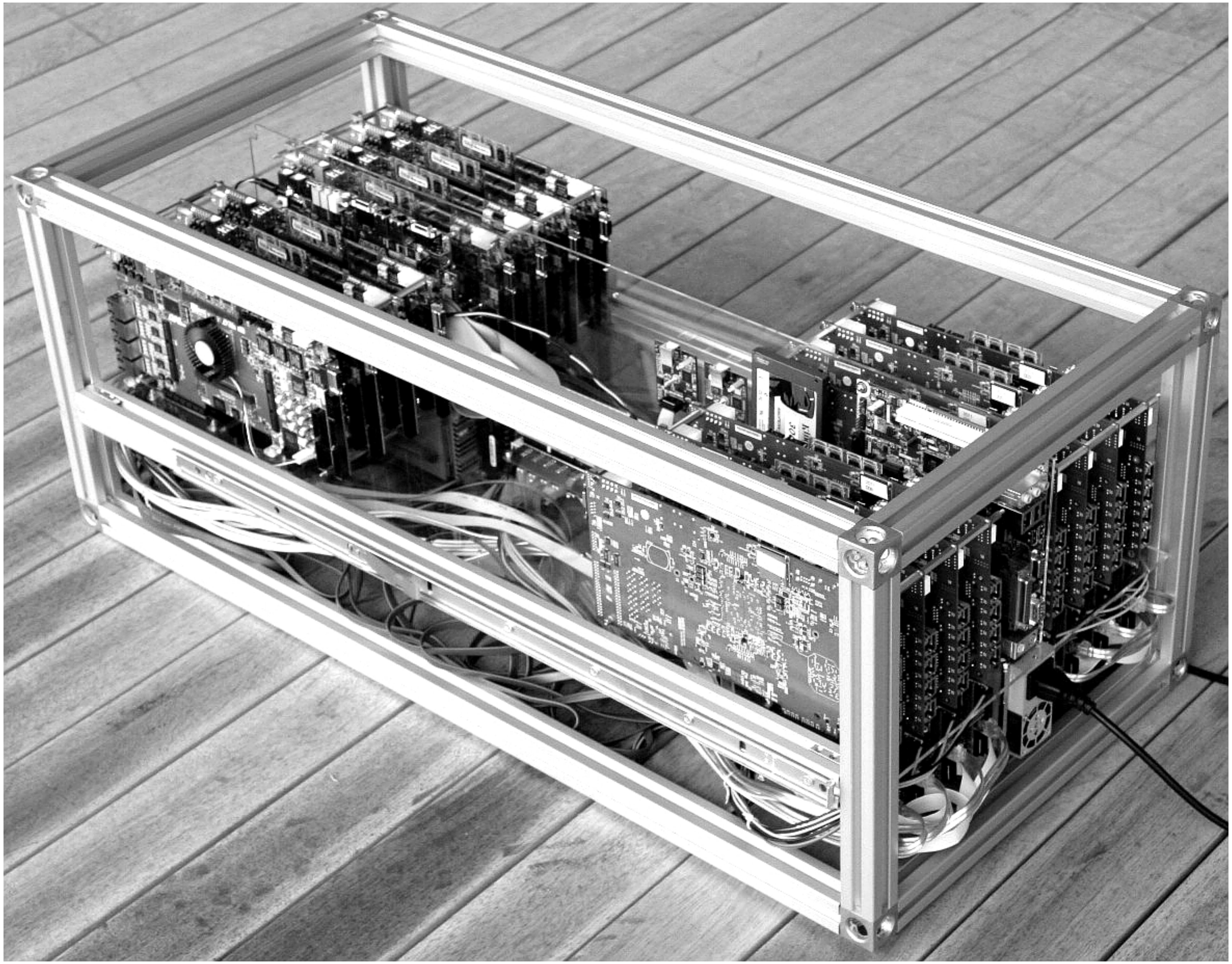
Chimera research questions

- Can information flow be derived from capabilities?
- How can we scale CHERI's capability model beyond a cache coherent CPU cluster?
- As cache coherency weakens, how can we support useful development and debugging models?
- What are efficient, scalable, and secure mappings of SCIEL computation into Chimera?

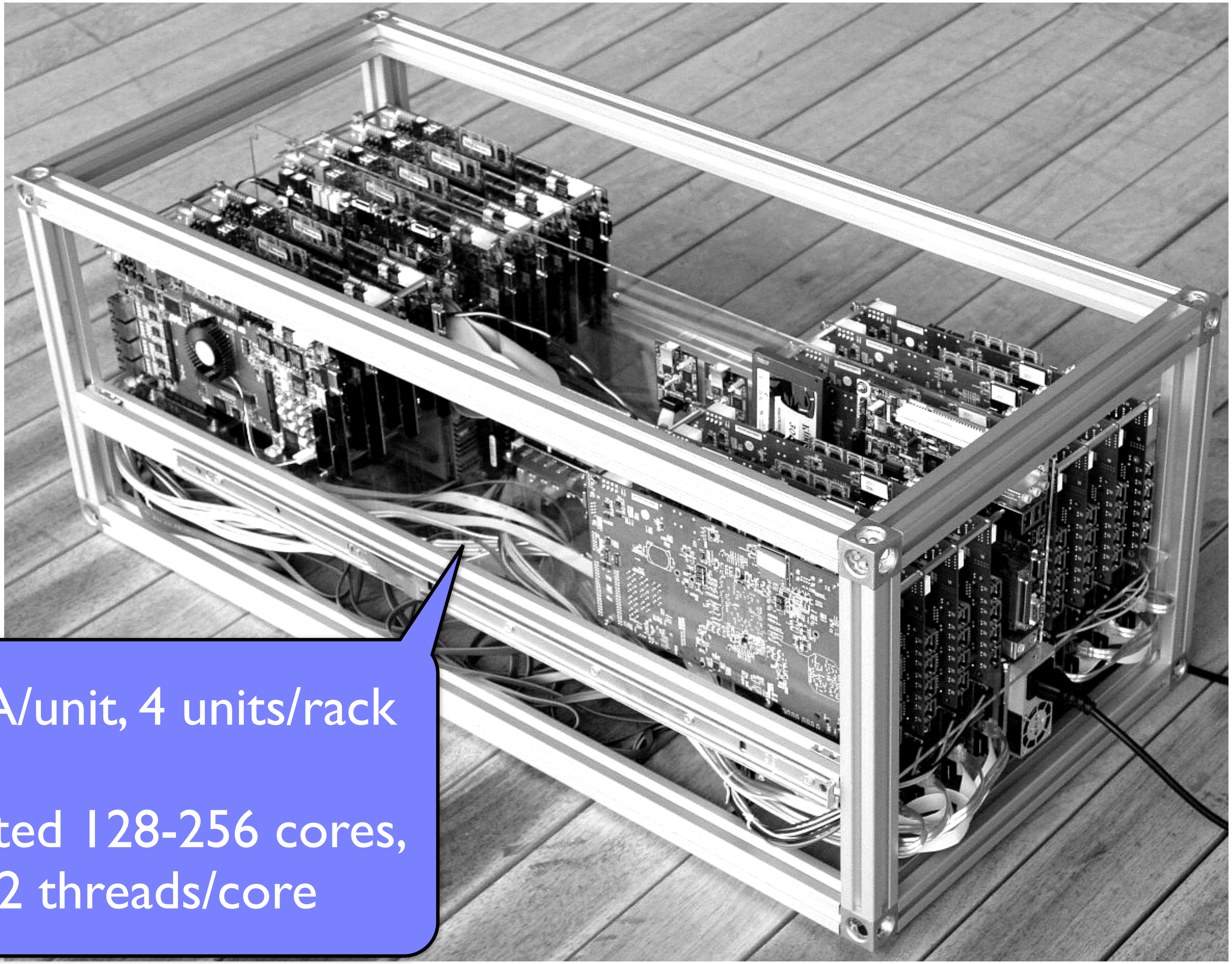
Chimera: Concept Implementation


heat sink
manycore layer with L1 caches
on-chip network and L2 caches
DRAM
DRAM
FLASH/MRAM
router and electro-optical interface
photonic modulators and receivers
link to another board, box or external interface
photonic and electrical PCB

Multi-FPGA Platform



Multi-FPGA Platform

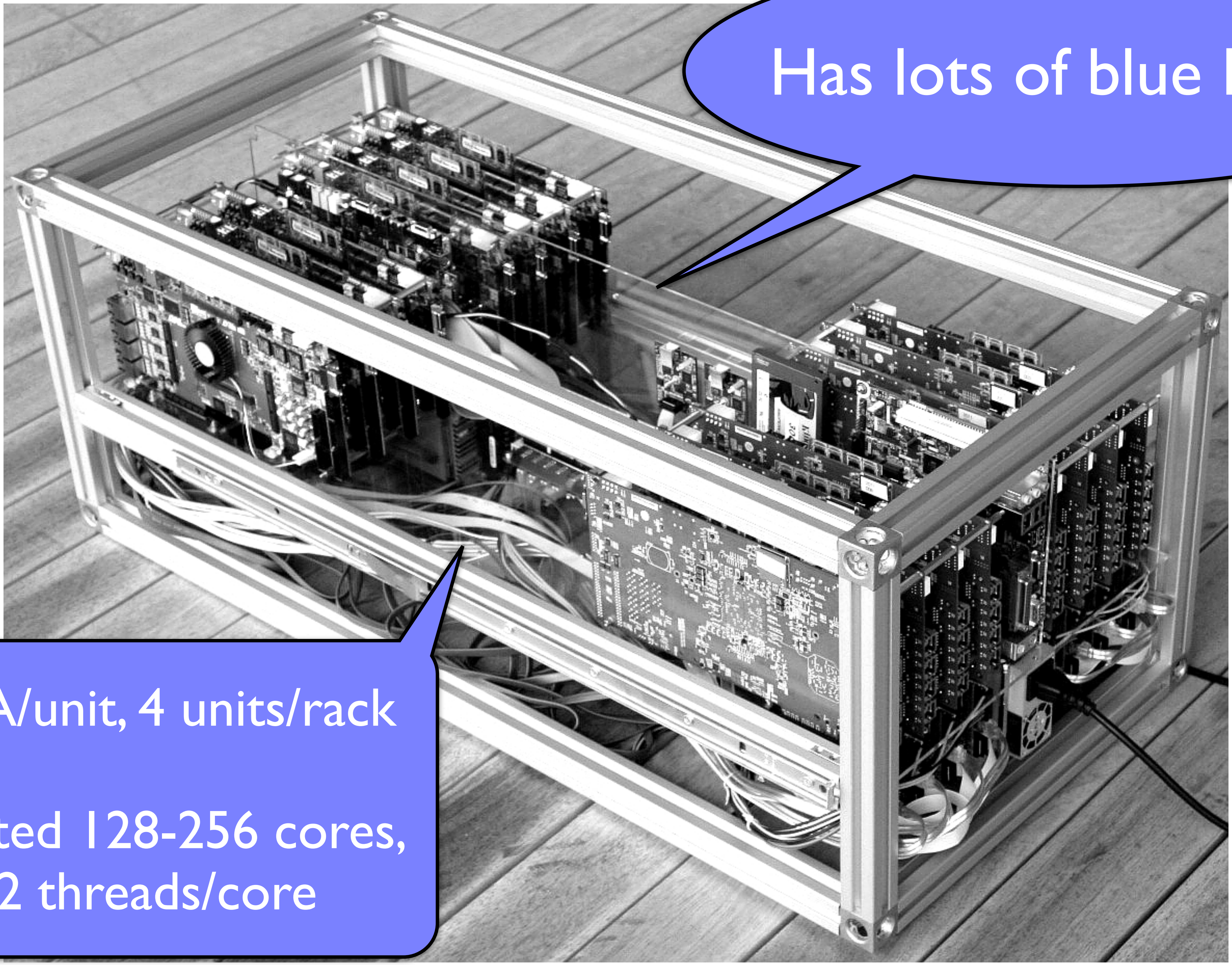


16 FPGA/unit, 4 units/rack

Anticipated 128-256 cores, 16-32 threads/core

UNIVERSITY OF CAMBRIDGE

Multi-FPGA Platform



Has lots of blue LEDs!

16 FPGA/unit, 4 units/rack

Anticipated 128-256 cores, 16-32 threads/core

Chimera progress

- Multithreaded CHERI2 prototype
  - can boot FreeBSD on one of the threads
- Multicore CHERI prototype
  - Prototype cache coherency scheme
- Multithreaded + multicore = fast simulation of 1000s of cores

RDSF

Resilient distributed switching fabric

- RDSF is a **high-dimensional structure** with a switchlet for every compute node
  - Multi-path redundant topology for performance, resilience and security
  - Components are functionally interchangeable
  - High(er) performance through closer processor-network affinity
  - Remap network topology to match program data flow

Non-traditional world view!

RDSF research questions

- Does data-centre mesh networking improve resilience, performance & power use?
- What are the interesting topologies to overlay over RDSF?
- How do we design a scalable network subsystem with low latency and high throughput?
- How do we architect a hierarchical control system that supports 10^6 switching elements?
- How can we implement efficient and resilient distributed accounting and audit for RDSF?
- What "alternative" semantics can we support: ordering, MPI, Chimera…?

RDSF progress
NetFPGA 10G infrastructure

- OpenFlow 1.0 Bluespec switch prototype with integrated CHERI2 processor
- More work on 10G platform and moving the community from the 1G platform

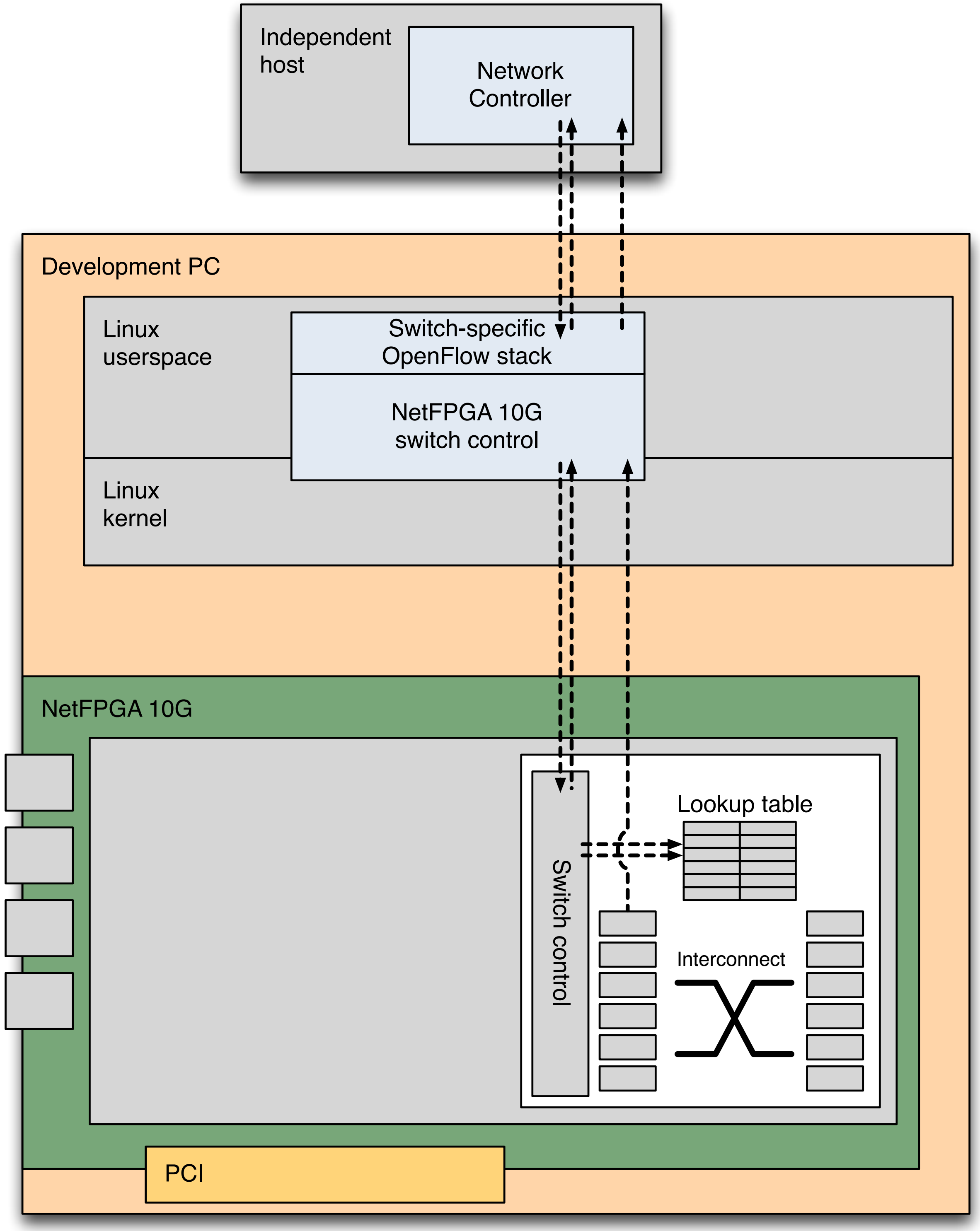
TPSC

Trustworthy programmable switch controllers

- Trustworthy platform for switch control
  - Platform for switch control applications
  - CHERI-based security model
- Distributed switch controllers
  - Integrated with RDSF: one-to-one with switchlets
  - Distributed control for resilience and performance
  - Scaling from one controller to one million?
- Formal grounding for both isolation and distribution properties

TPSC

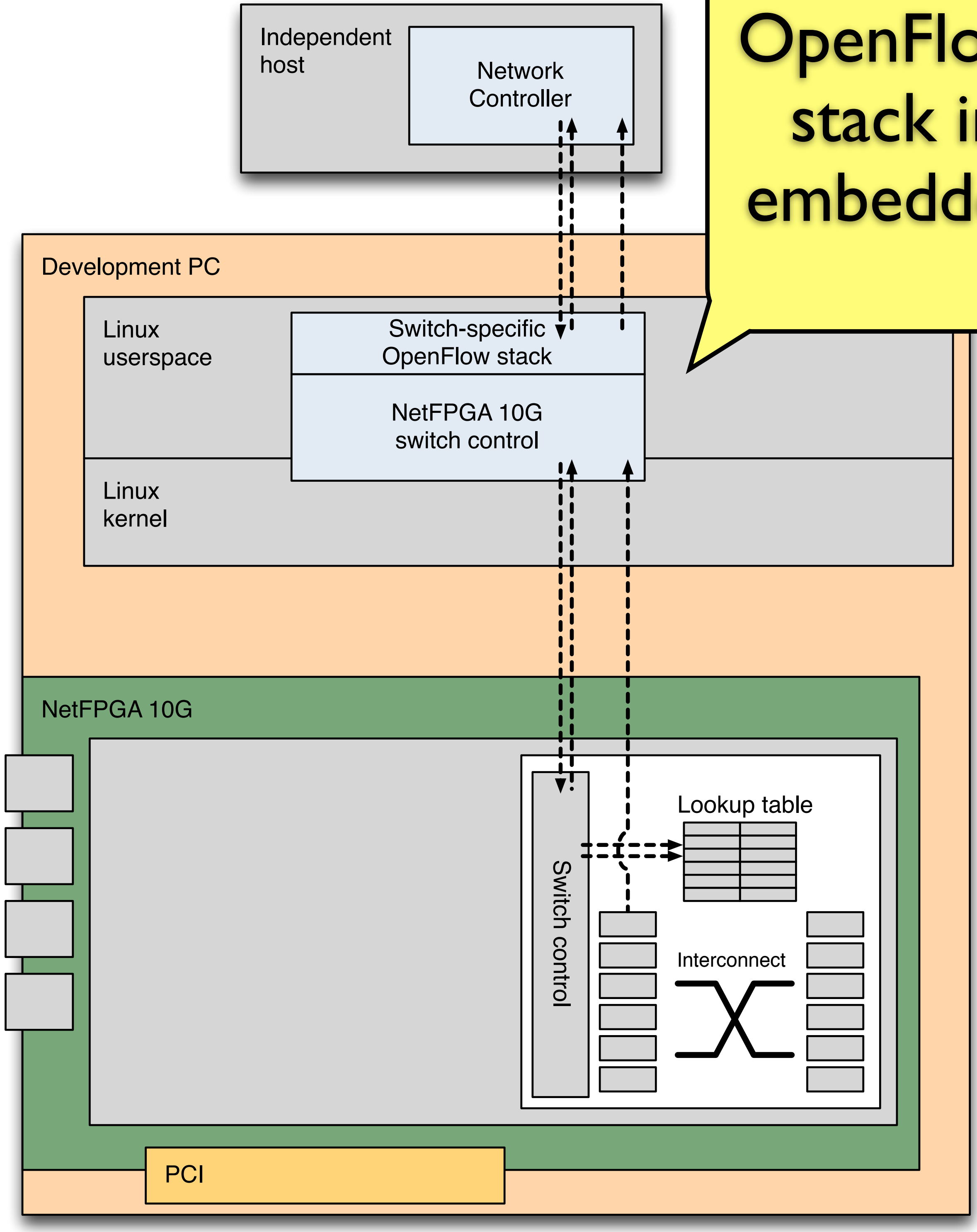
Trustworthy Switches and Switch Controllers



- Develop trustworthy switch and switch controller platforms for use in RDSF.
- Integrate verified CHERI processor and software stack with switch control
- Develop and integrate verified Bluespec switch core
- Both distribute and compartmentalise switch control to improve resilience and security

TPSC

Trustworthy Sw



Conventional switches run OpenFlow (or similar) management stack in Linux or BSD on COTS embedded processor, isolated from processing path.

witch

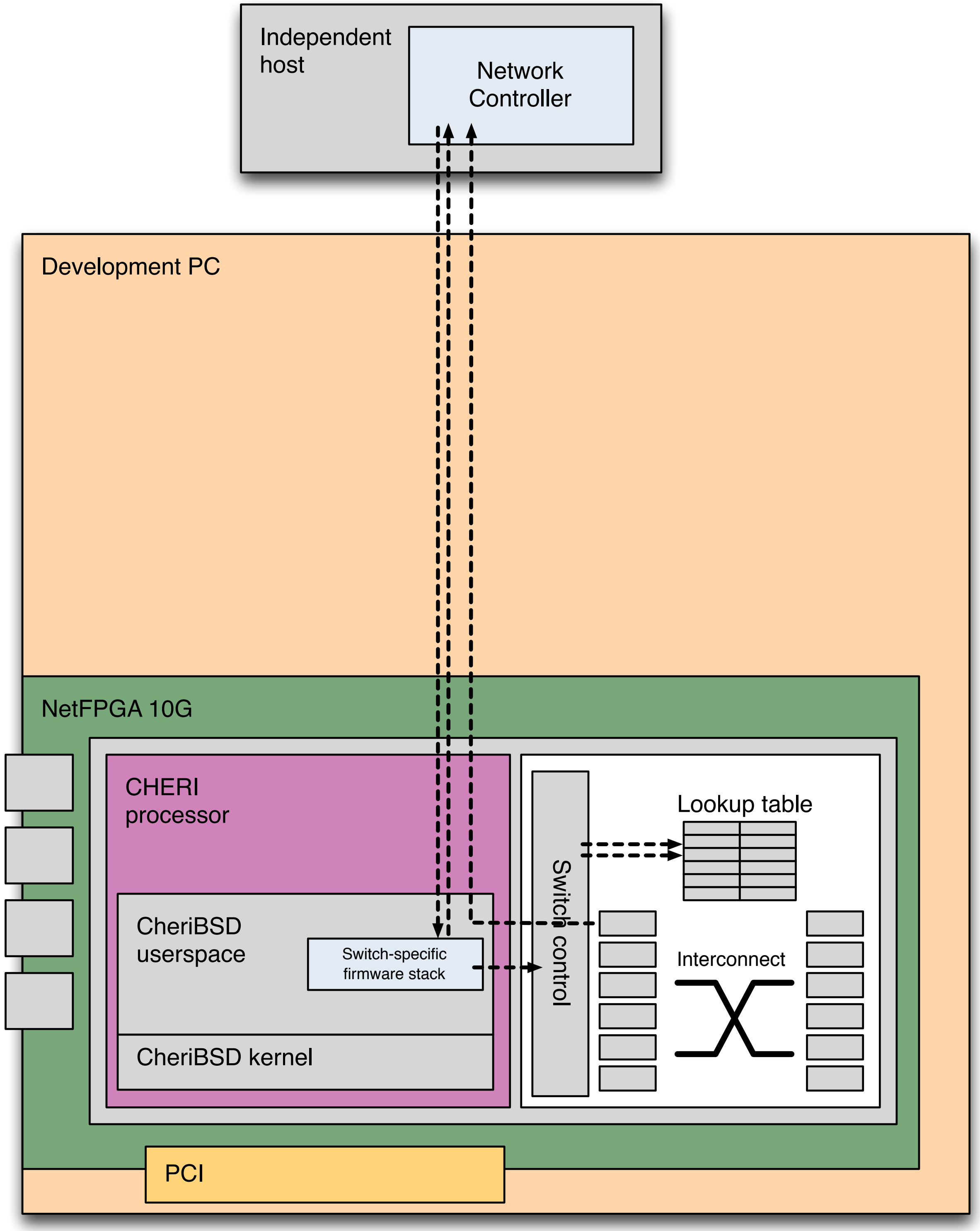
SF.

- Integrate verified CHERI processor and software stack with switch control
- Develop and integrate verified Bluespec switch core
- Both distribute and compartmentalise switch control to improve resilience and security

UNIVERSITY OF CAMBRIDGE

TPSC

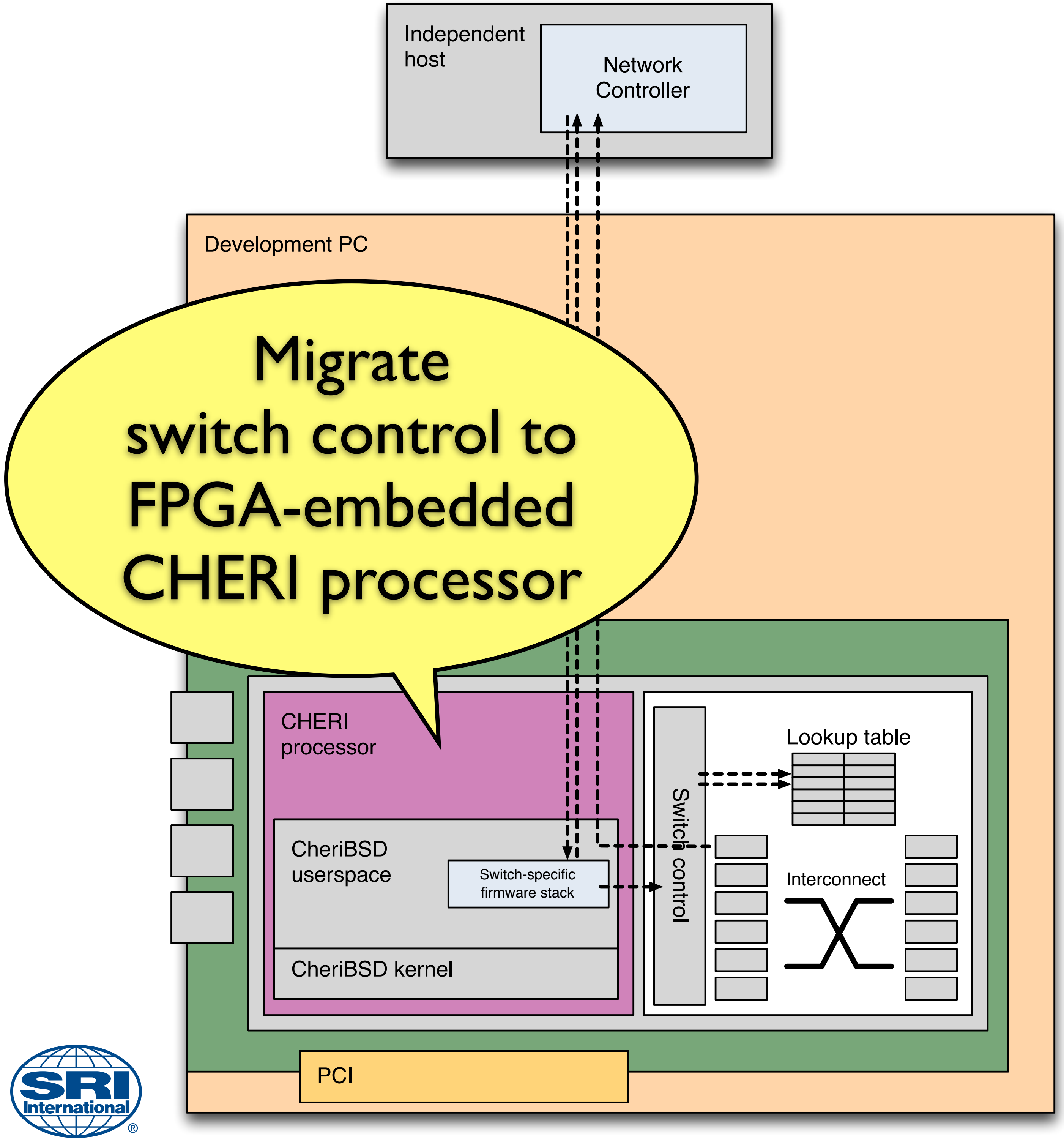
Trustworthy Switches and Switch Controllers



- Develop trustworthy switch and switch controller platforms for use in RDSF.
- Integrate verified CHERI processor and software stack with switch control
- Develop and integrate verified Bluespec switch core
- Both distribute and compartmentalise switch control to improve resilience and security

TPSC

Trustworthy Switches and Switch Controllers

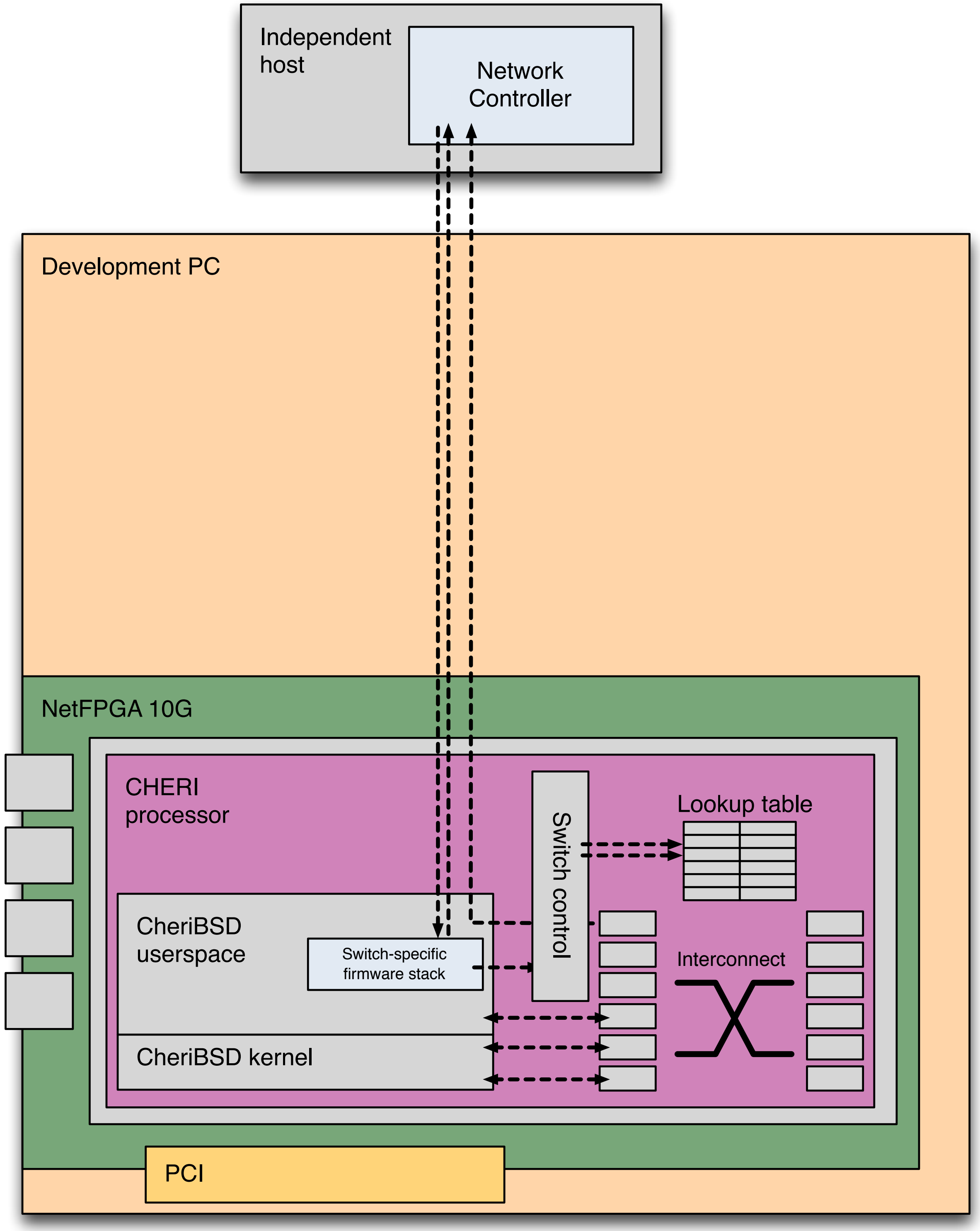


Migrate switch control to FPGA-embedded CHERI processor

- Develop trustworthy switch and switch controller platforms for use in RDSF.
- Integrate verified CHERI processor and software stack with switch control
- Develop and integrate verified Bluespec switch core
- Both distribute and compartmentalise switch control to improve resilience and security

TPSC

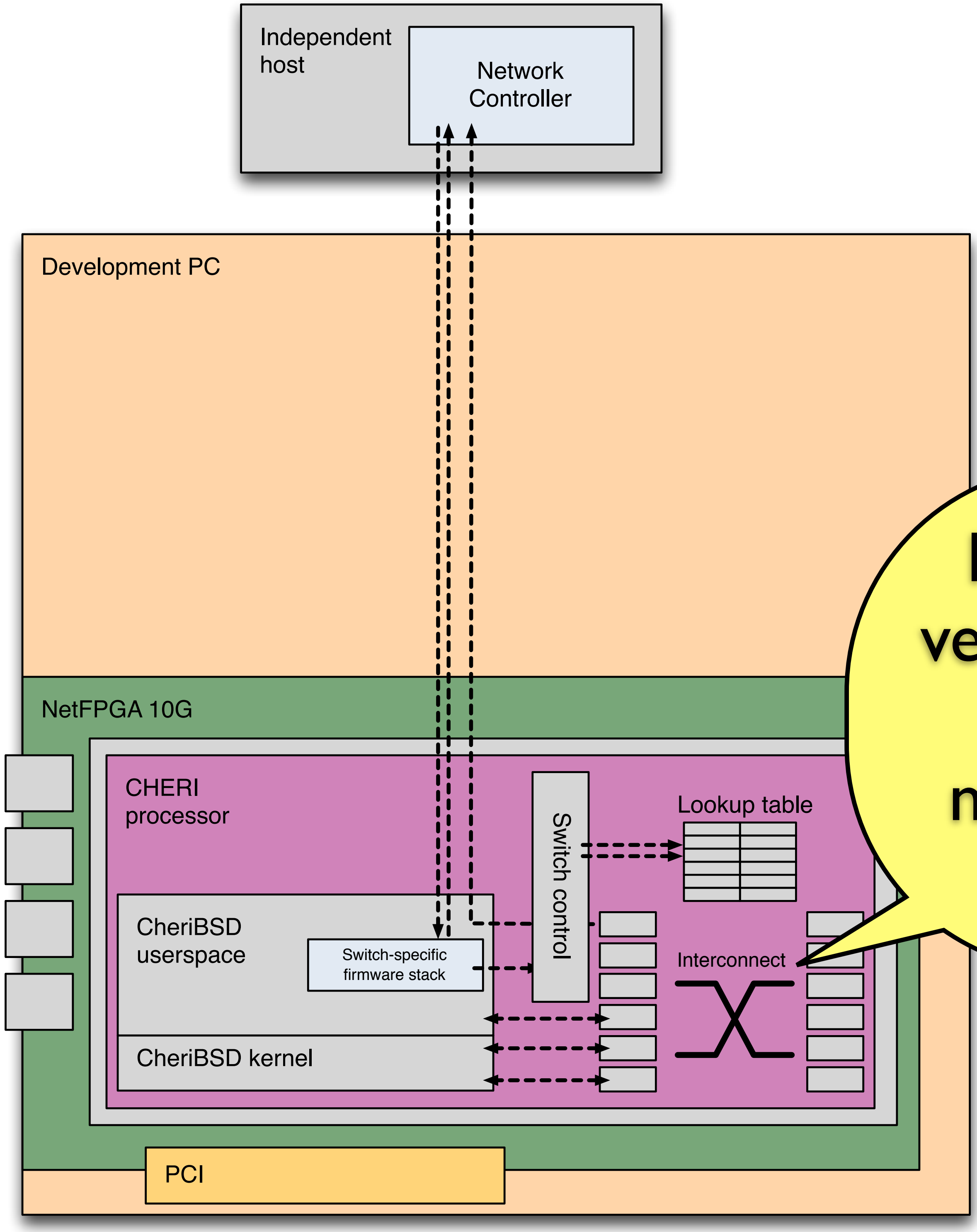
Trustworthy Switches and Switch Controllers



- Develop trustworthy switch and switch controller platforms for use in RDSF.
- Integrate verified CHERI processor and software stack with switch control
- Develop and integrate verified Bluespec switch core
- Both distribute and compartmentalise switch control to improve resilience and security

TPSC

Trustworthy Switches and Switch Controllers



- Develop trustworthy switch and switch controller platforms for use in RDSF.
- Integrate verified CHERI processor and software

e

control to improve resilience and security

Integrate CHERI and formally verified Bluespec switch pipeline, allowing compartmentalised network processing inline with packet paths

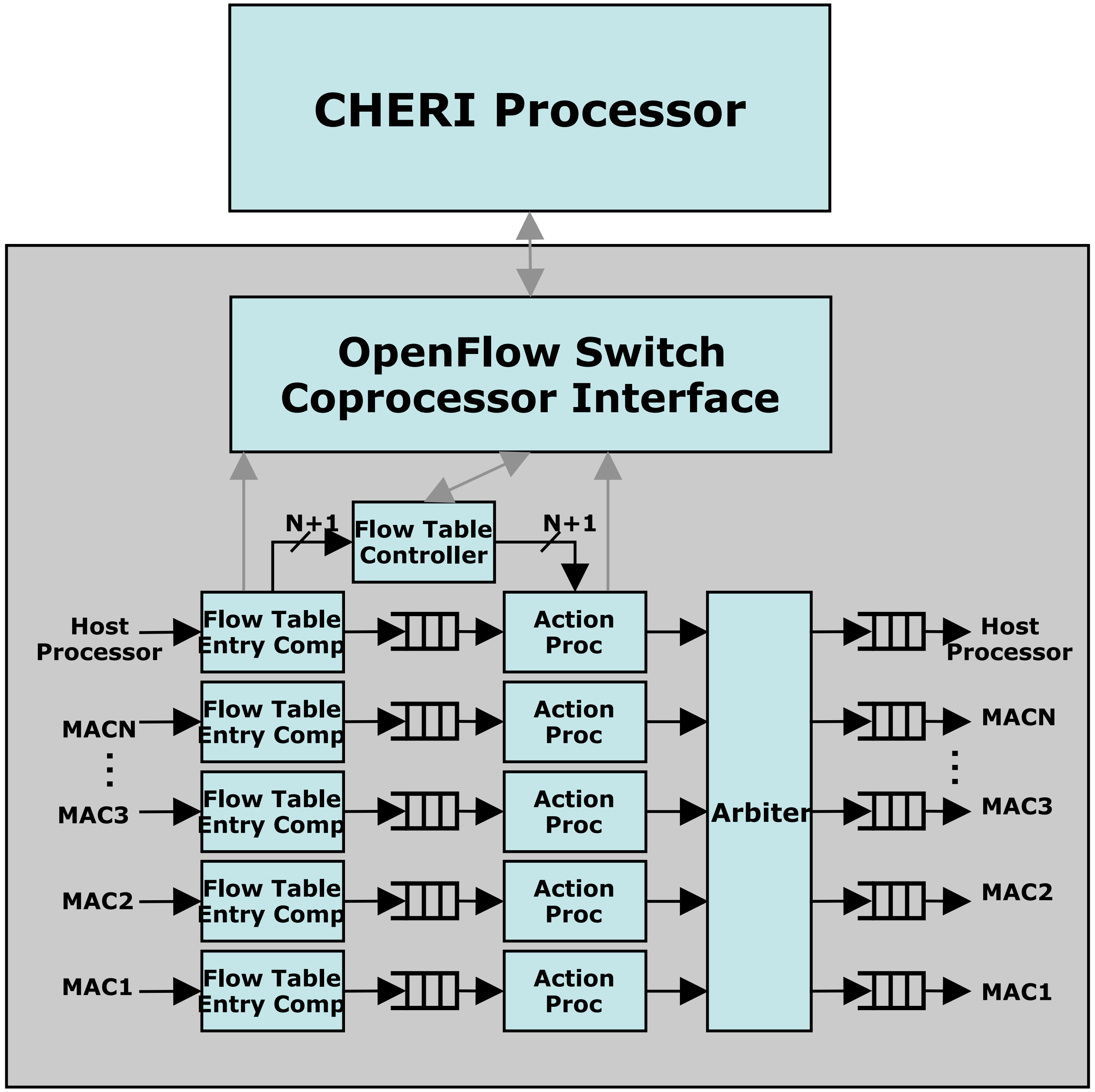
TPSC research questions

- How can we apply CHERI-based compartmentalisation to switch and switch controller platforms to improve security?
- Can Bluespec-PVS links being developed in CRASH be used to develop a formally verified switching path?
- Can Secure SDN scale to large numbers of distributed switch controllers to improve resilience?

TPSC - First Cut

- DE4
- BeriBSD
- Open vSwitch on CHERI
  - Pure software switch with NetMap
  - Goal to use CHERI protection features

BSV OpenFlow Switch



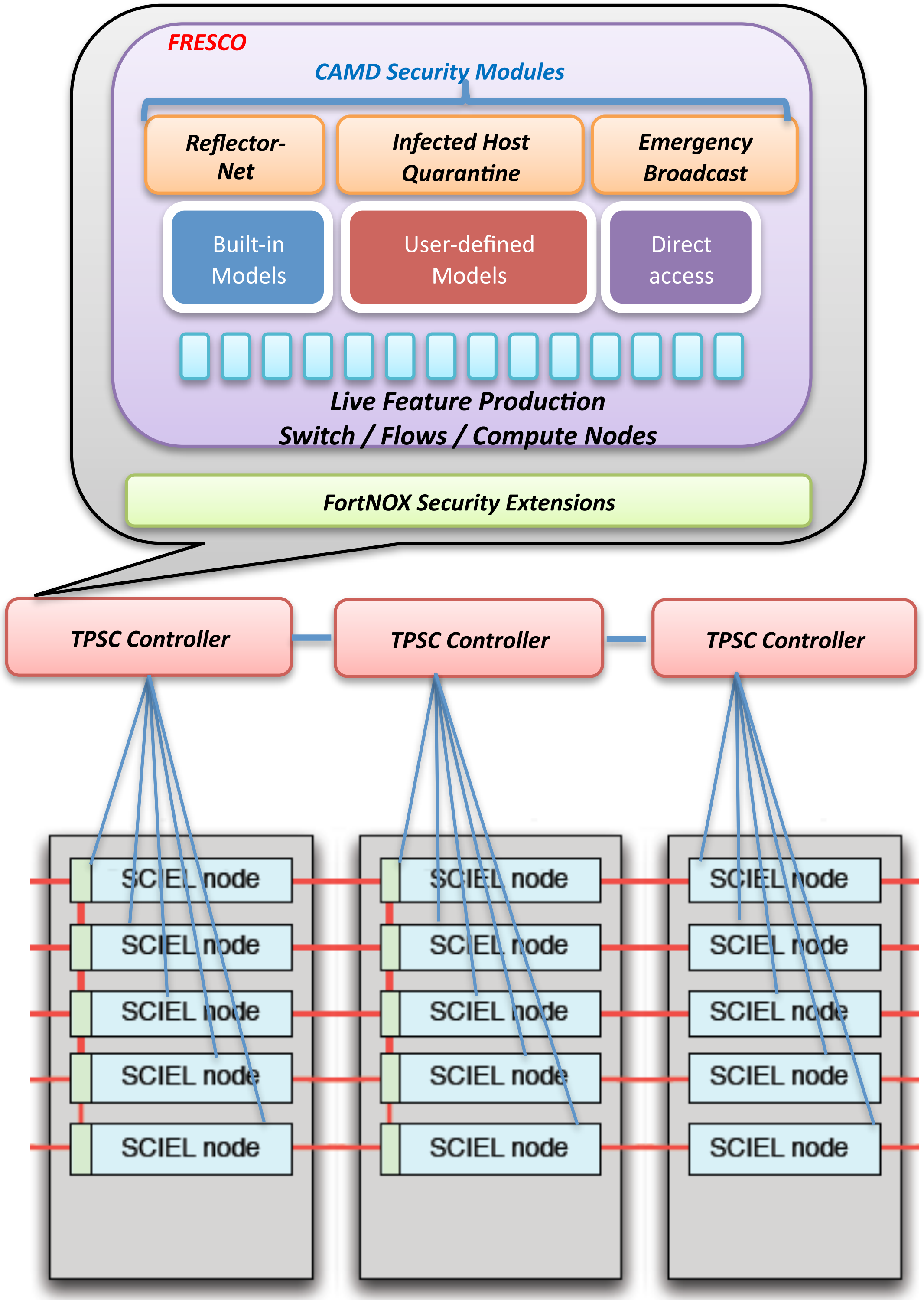
- 10Gbps HW switch design on NetFPGA-10G
- OpenFlow v1.0.0
- 2400 Lines of modular/ parameterized Bluespec
- ~50ns latency to on-FPGA CHERI Processor

TPSC Directions

- Hardware-software architecture developed, based on CHERI and RDSF designs
- Continue to develop Bluespec-PVS formal methods bridge in CRASH, to apply to TPSC switch elements

CAMD

Cloud analysis and misuse detection

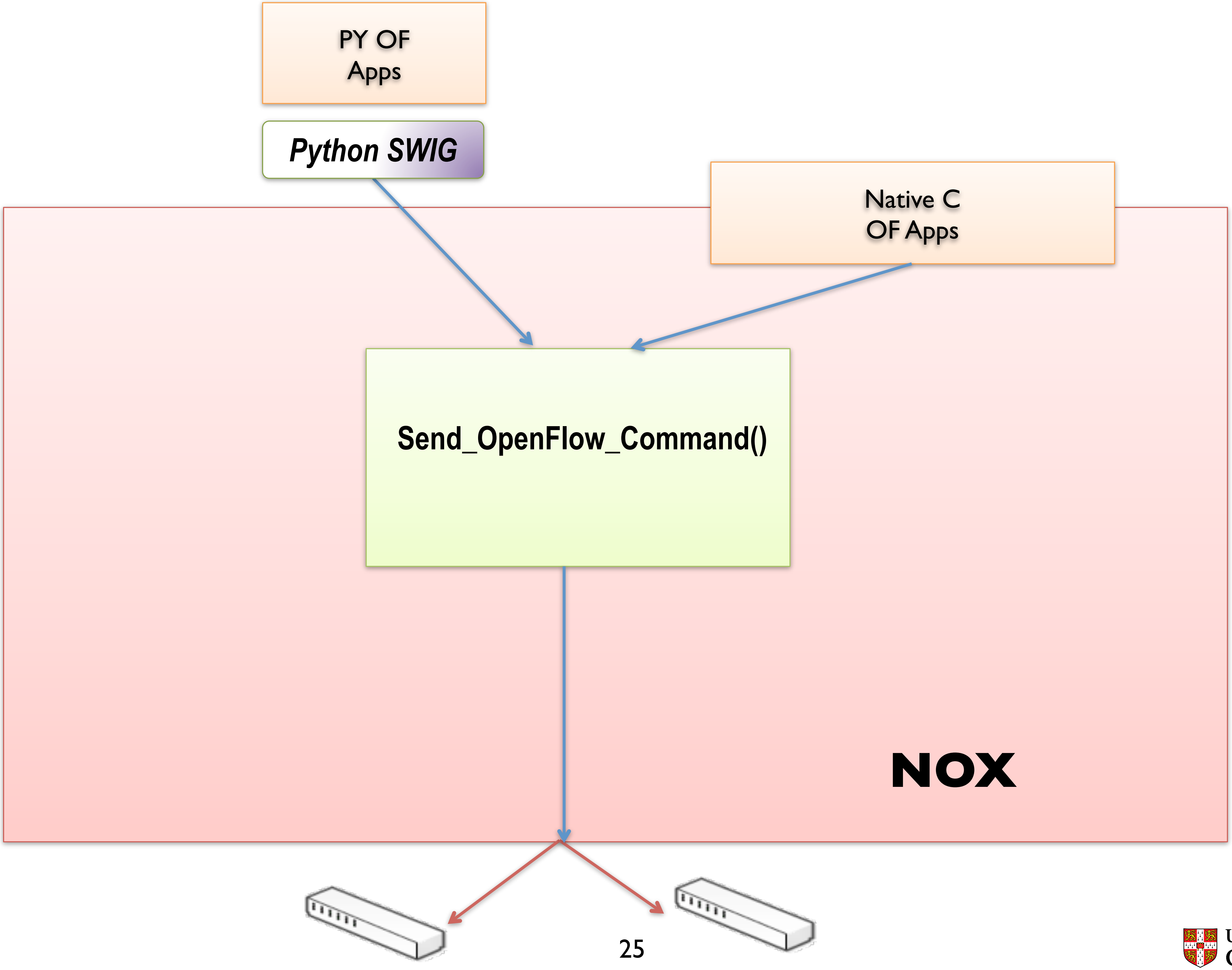


- Existing approaches rely on centralised visibility and control
    - e.g., OpenFlow switches/ controllers
- Distribute both switch management functionality and analytics/misuse detection/ remediation
- CAMD offered by and for multiple tenants, as well as data center owners
    - e.g., APIs for IDSs, reflector nets, emergency broadcasts, IPSs, BotHunter, ...

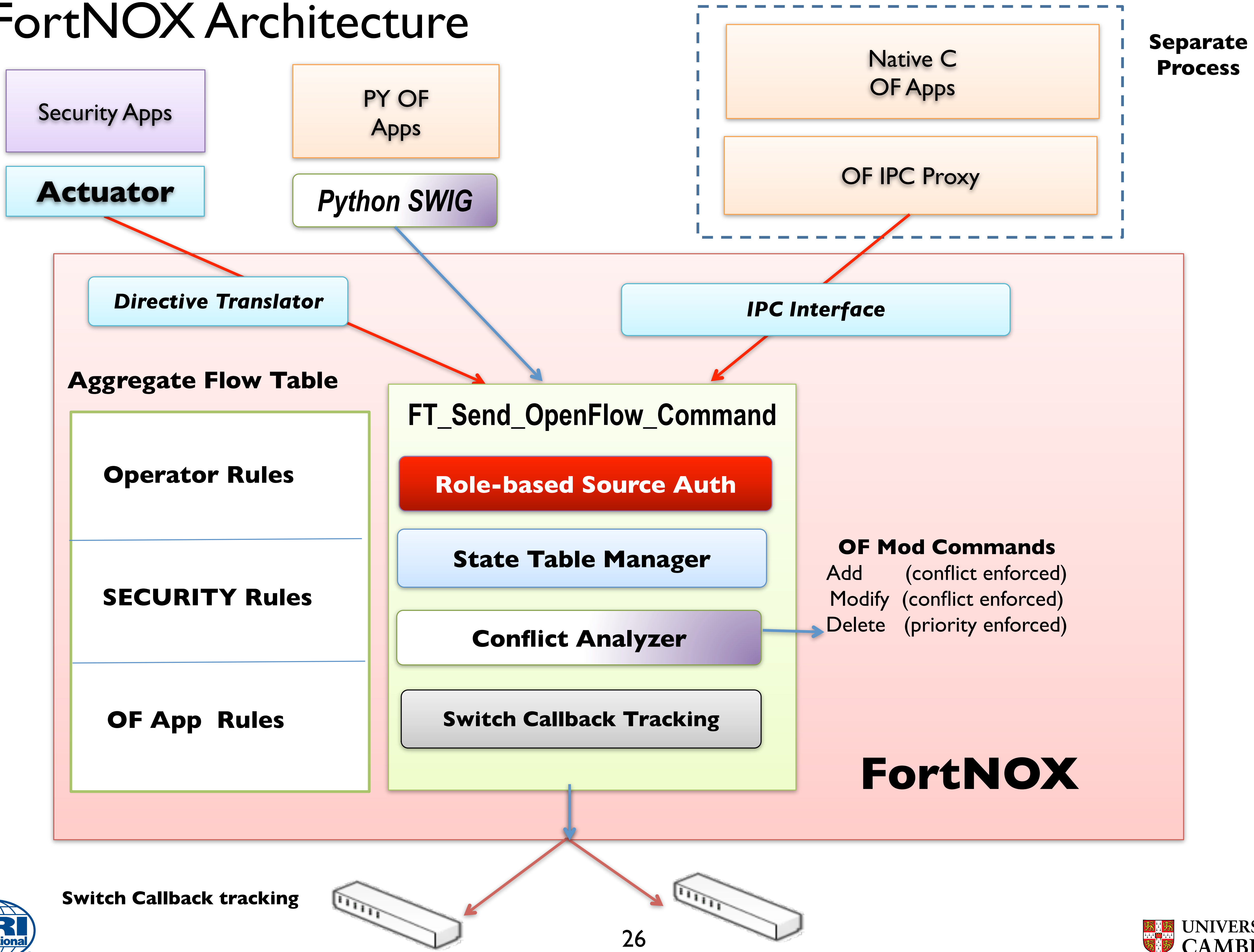
CAMD research questions

- How can we overcome inherent synchronisation and distributed enforcement problems?
- How can we achieve prompt, efficient, accurate, and conflict-free distributed accounting, policy changes, garbage collection, and resilience despite would-be adversities?
- How can we automatically detect and mitigate vulnerabilities in SSDN applications?
- What adversary models can be accommodated by CAMD in MRC2?

Classic NOX Architecture



FortNOX Architecture



Separate Process

Native C OF Apps

OF IPC Proxy

Security Apps

Actuator

PY OF Apps

Python SWIG

Directive Translator

IPC Interface

Aggregate Flow Table

Operator Rules

SECURITY Rules

OF App Rules

FT_Send_OpenFlow_Command

Role-based Source Auth

State Table Manager

Conflict Analyzer

Switch Callback Tracking

OF Mod Commands

Add (conflict enforced)

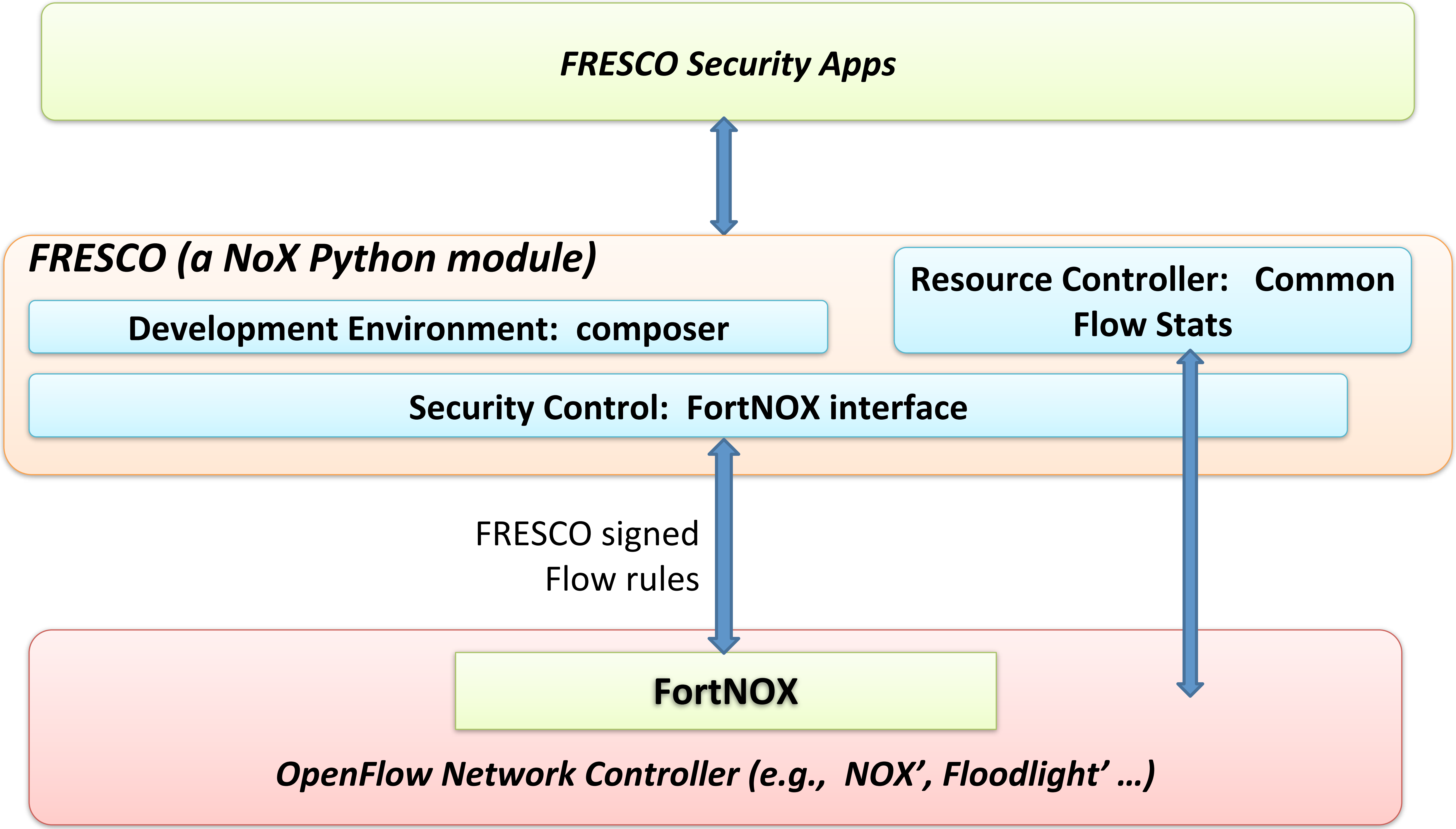
Modify (conflict enforced)

Delete (priority enforced)

FortNOX

Switch Callback tracking

FRESCO



FRESCO Security Apps

FRESCO (a NoX Python module)

Development Environment: composer

Resource Controller: Common Flow Stats

Security Control: FortNOX interface

FRESCO signed Flow rules

FortNOX

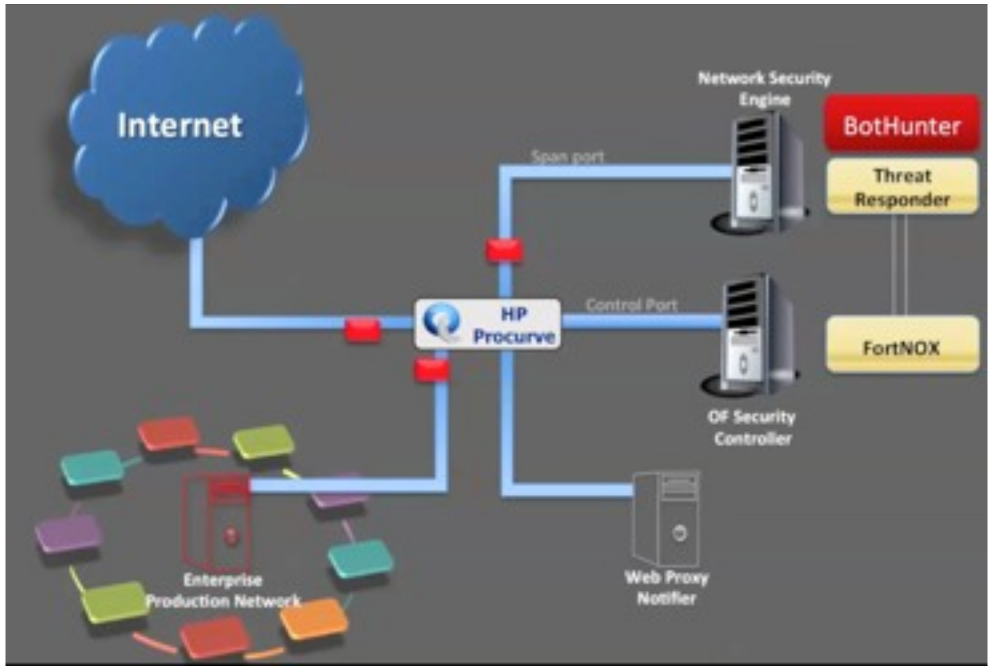
OpenFlow Network Controller (e.g., NOX', Floodlight' …)

CAMD Progress

- Published the Fresco Application Framework for integrating security application into an Secure SDN control stack
- Designing new security extensions for switches to facilitate high performance threat mitigation services for cloud computing
- Produced performance and expressibility evaluations for our FRESCO Security scripting language and demonstrated the linking of legacy security services to the FRESCO mediation services
- Working toward an evaluation of a software implementation of our security extensions within an OpenFlow switch, and developing attack/response demonstrations to illustrate our design

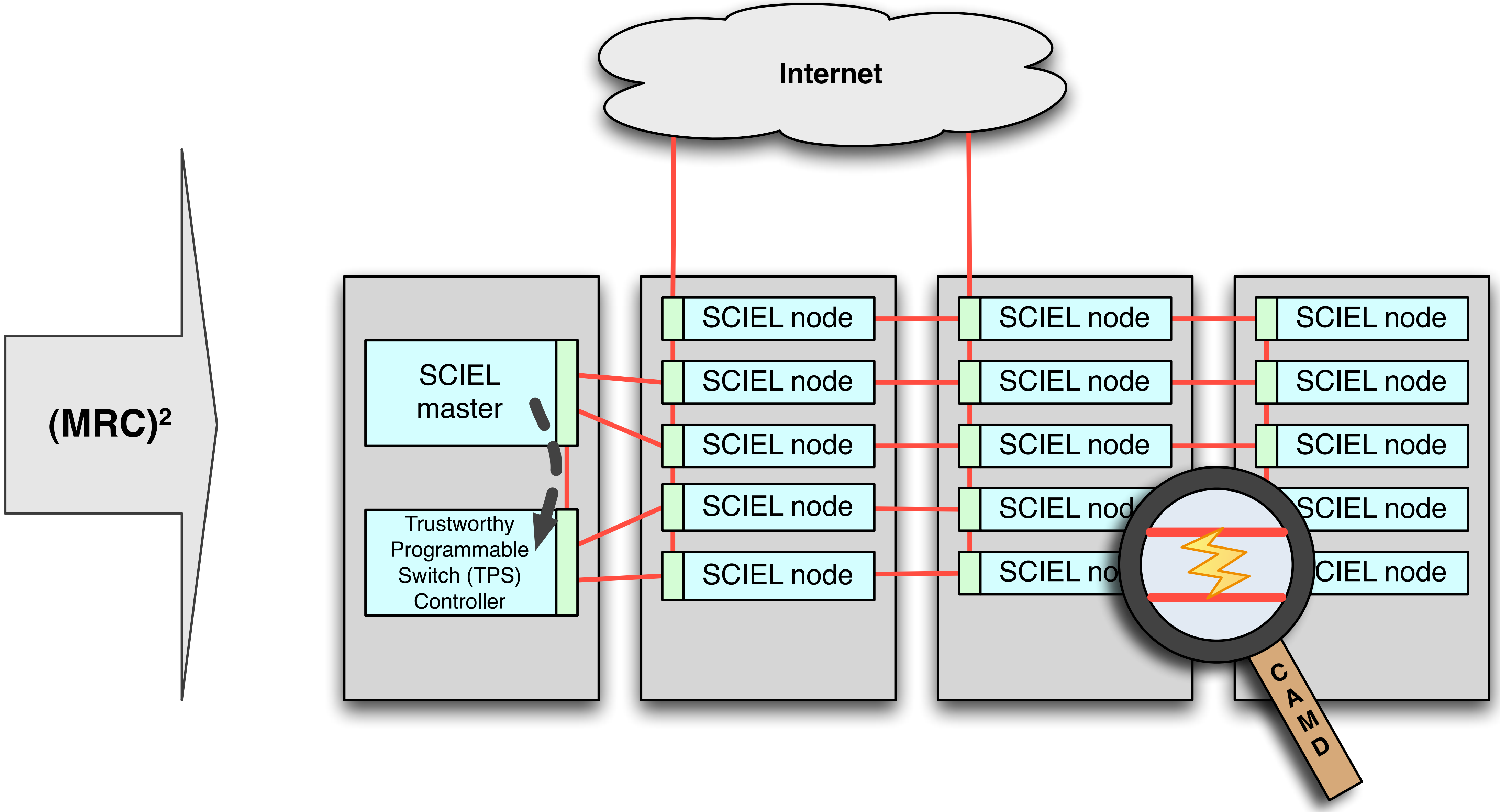
CAMD Demos

- Demo 1: Security Constraints Enforcement
  - A Demonstration of policy enforcement using an OpenFlows
  - Security Mediation Service
- Demo 2: Reflector Nets
  - OpenFlow Security App: Demonstrating dynamic attack redirection
- Demo 3: Automated Quarantine
  - OpenFlow Security App: Demonstrating infection quarantine of a malware infected local host



SCIEL

A programming framework for secure clouds

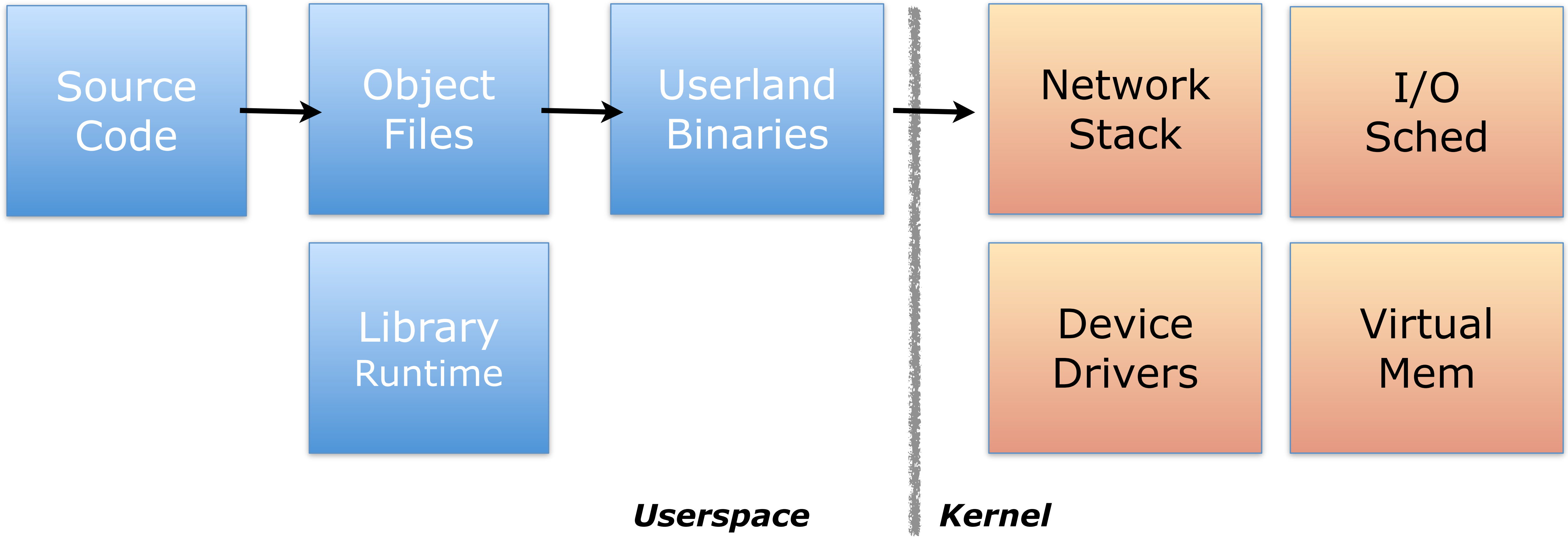


SRI International | UNIVERSITY OF CAMBRIDGE

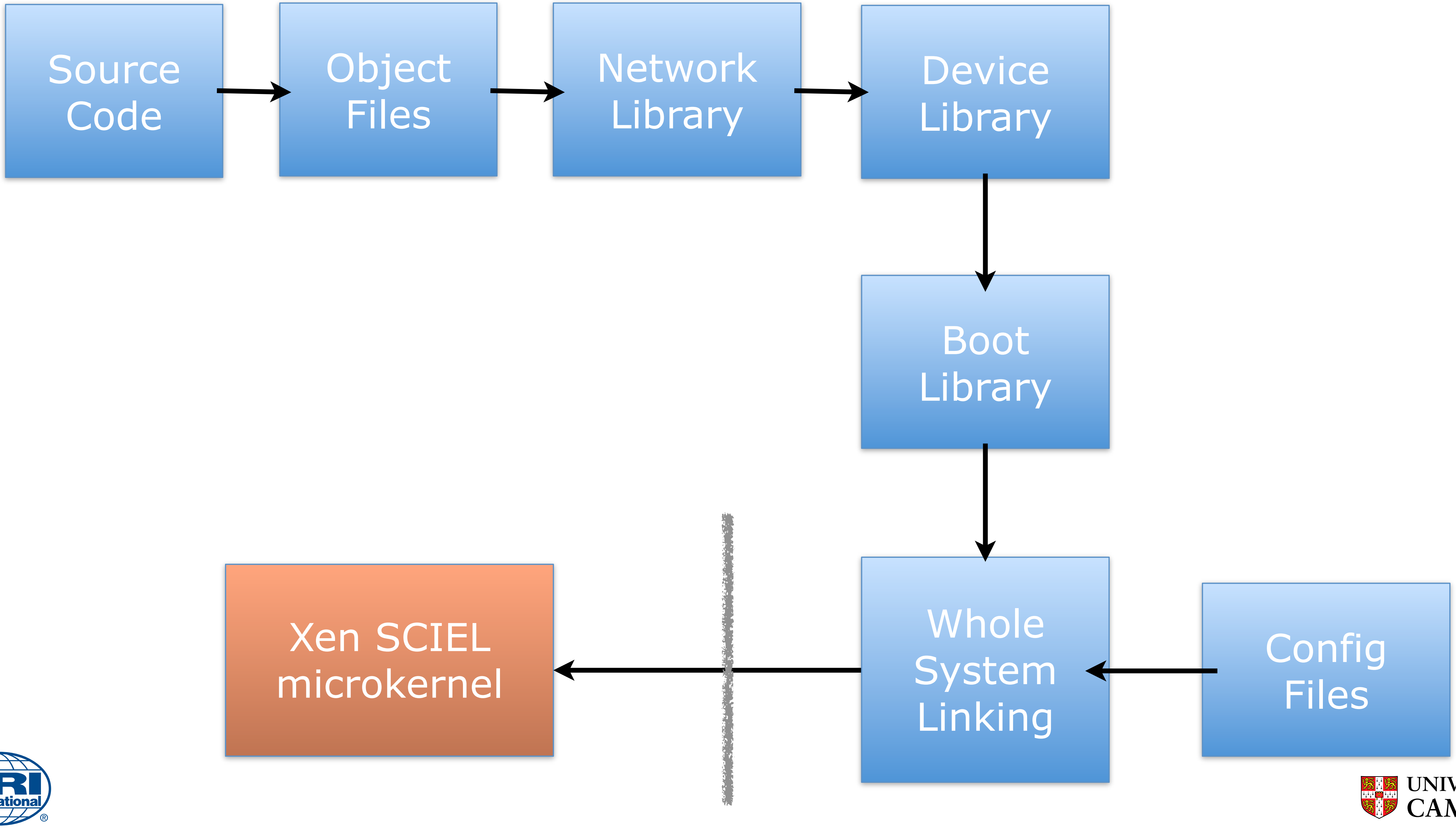
SCIEL Concepts

- SCIEL - a programming model for distributed computation across a multi-tenant cloud fabric
- A thin hypervisor layer (e.g. Xen or CheriBSD) executes a task-parallel computation across a cluster
- SCIEL computation is fault-tolerant
  - individual nodes can be specialized to the computation environment

SCIEL nodes: FreeBSD



SCIEL nodes: specialised



SCIEL Unikernel Image Size

Appliance	Unikernel image size
DNS	0.184 MB
Web Server	0.172 MB
Openflow learning switch	0.164 MB
Openflow controller	0.168 MB

SCIEL Status

- Released alpha of the Mirage exokernel OS that can run on the public cloud
  - www.openmirage.org
- Developed a FreeBSD/CheriBSD kernel module version of Mirage
  - enables apples-for-apples comparison of Mirage vs. conventional designs, e.g. network stack

Example of Cross-Cutting Work

SDNsim

SDNsim

- SDNsim - a Software Defined Network macro-simulator framework
- Provides an abstraction layer to replicate behavior of network nodes
- Specification → emulation + simulation
  - Simulation: NS3 network simulator platform
  - Emulation: Xen Cloud platform

OpenFlow Challenges

- Control Centralization
  - Single point of failure
  - Controller becomes the performance bottleneck
  - Control granularity: recipe is unclear
- Control distribution
  - Hard to get it right, e.g. routing protocols
  - Contradicting goals: latency vs scalability

SRI International®

UNIVERSITY OF
CAMBRIDGE

Experimenting with OpenFlow in Mirage on Xen

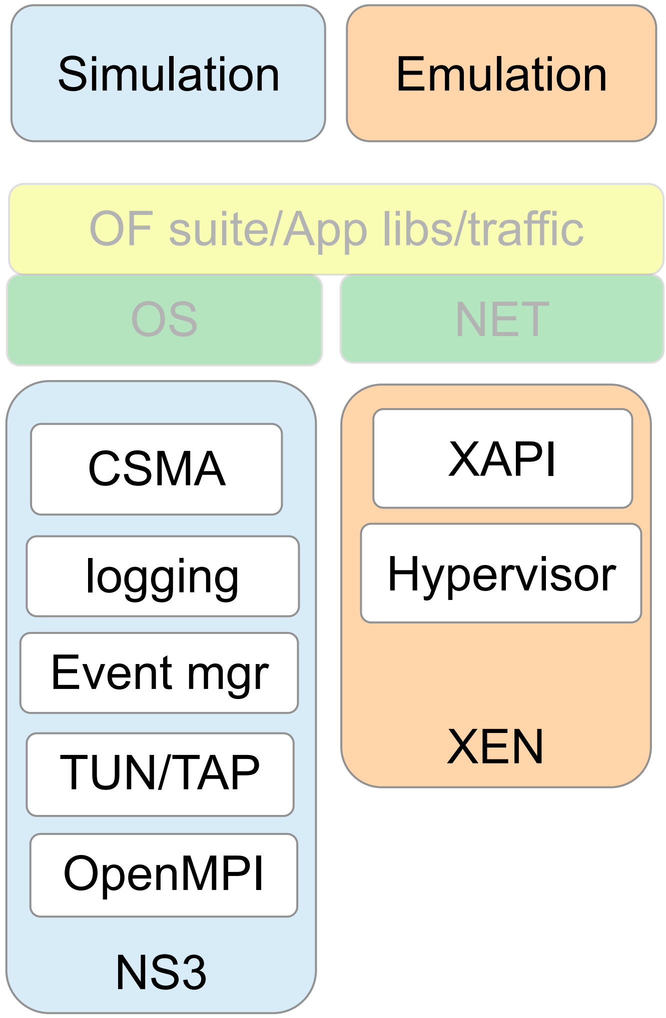
Controller	*Throughput(kreq/sec)*		*Latency (kreq/sec)*	
	avg	*std*	*avg*	*std*
Nox destiny fast	122.6	4.9	27.5	0.1
Maestro (Java)	13	0.9	26.4	1.0
Mirage UNIX	79.2	0.5	24	0.2
Mirage Xen	97.6	1.5	23.9	0.9

- Cbench: generate OF packet_in packets and measure throughput of controller
- Parameters: 16 switches, 100 mac/switch, single thread, 20 runs

SRI International

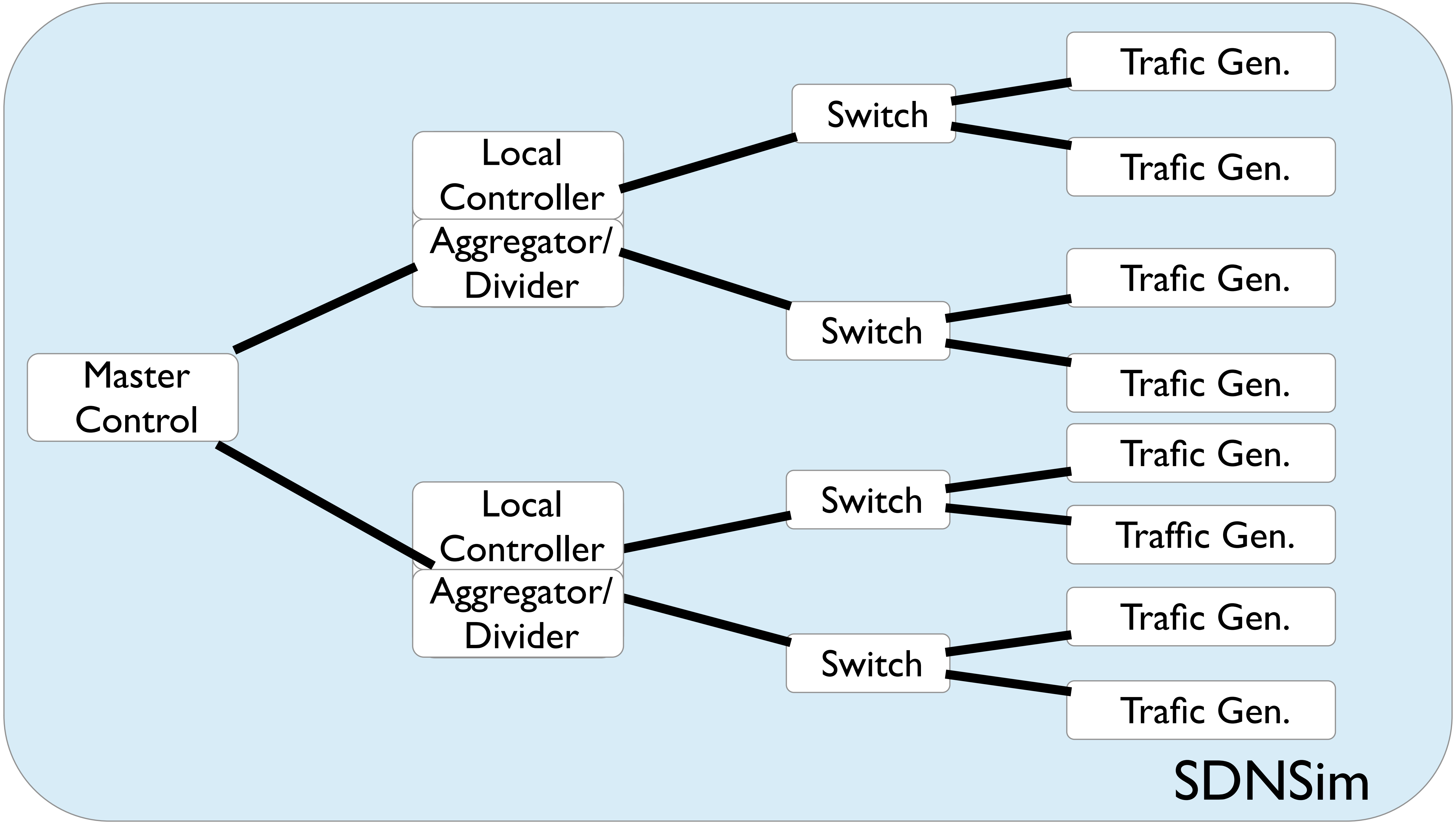
UNIVERSITY OF CAMBRIDGE

SDNsim Backends

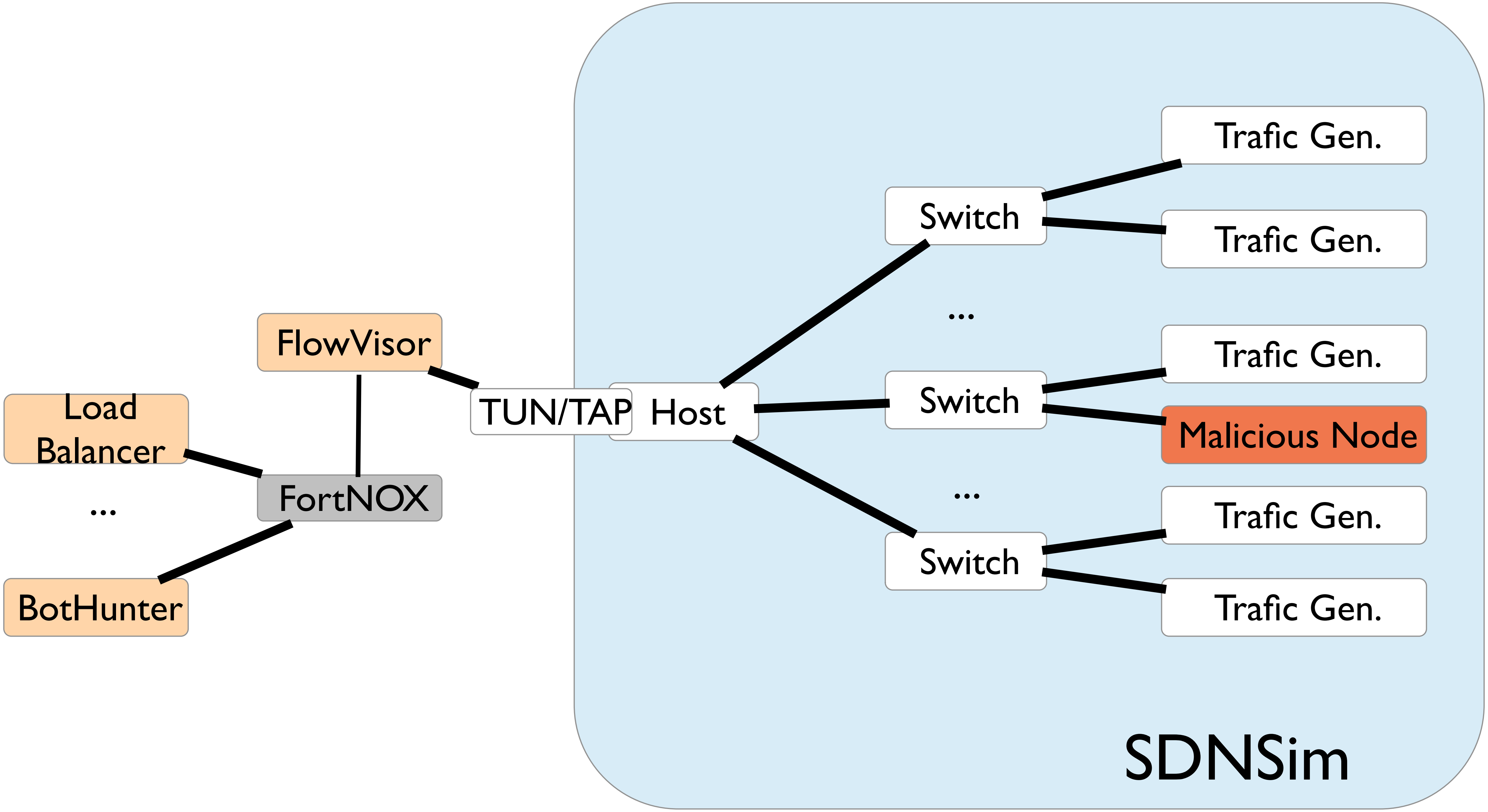


- Topology and node functionality described through a configuration file.
- Simulation: NS3
  - High precision event-driven
  - Many libraries to emulate links, distribute processing and communicate with external entities.
- Emulation: XEN Cloud Platform
  - MiniOS-based bootable kernels.
  - Xen API provides accurate resource provisioning.
  - Tunable clock rate to reduce XEN processing noise.

Use Case: RDSF



Use Case: FRESCO



Conclusion

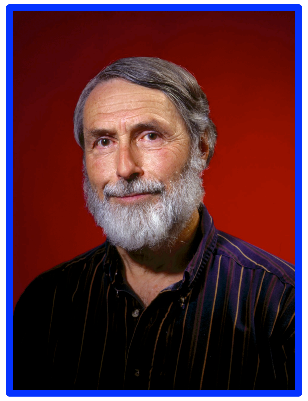
- Infrastructure development coming on a pace:
  - Multi-core and multithreaded capability processor and coherent memory subsystem in prototype
  - Programming models for the cloud
    - Mirage (for SCIEL nodes) paper to appear soon
    - Enabled SDNsim hooked into FRESCO
  - FRESCO security application framework in NDSS 2013
  - 10GB/s FPGA-based switch infrastructure (FPGA hardware and interface logic as part of NetFPGA10G project)
  - Prototype OpenFlow switch in Bluespec on NetFPGA10G

In the news

- New York Times article & video: Profiles in Science Peter G. Neumann: Rethinking the Computer at 80
  http://www.nytimes.com/2012/10/30/science/rethinking-the-computer-at-80.html
- Queue Portrait: Robert Watson

  http://queue.acm.org/detail_video.cfm?id=2382552

The (MRC)^2 team


Dr Peter G. Neumann

Dr Robert N.M. Watson

Dr Simon W. Moore

Dr Nirav Dave

Mr Brooks Davis

Dr Hassen Saidi

Dr Patrick Lincoln

Mr Phillip Porras

Mr Stacey Son

🟦 SRI International

🟥 University of Cambridge

MRC2 project members unable to attend the PI meeting:

Jonathan Anderson, David Chisnall, Matthew P. Grosvenor, Khilan Gudka, Asif Khan, Myron King, Anil Madhavapeddy, Andrew Moore Alan Mujumdar, Steven J. Murdoch, Robert Norton, Muhammad Shahbaz, Richard Uhler, Jonathan Woodruff, Vinod Yegneswaran, Dongting Yu

Backup Slides

Resilience

- CAMD - using software defined networks to reconfigure around faults
- RDSF - resilience through distribution of the switch fabric
- SCIEL - resilience by automating retries and allowing work to be redistributed to tolerate failed nodes
- Chimera - capability-based (CHERI) processors provide resilience through sandboxing