The OptoHPC simulator: Bringing OptoBoards to HPC-scale environments

Pavlos Maniotis, *Nikos Terzenidis, Nikos Pleros*
Aristotle University of Thessaloniki (AUTH), Greece

OMNeT++ Community Summit 2016
15 September 2016, Brno, Czech Republic

Outline

- Introduction
- The OptoHPC simulator architecture
- An OptoHPC use case: comparison performance analysis using the OptoHPC
- Conclusion

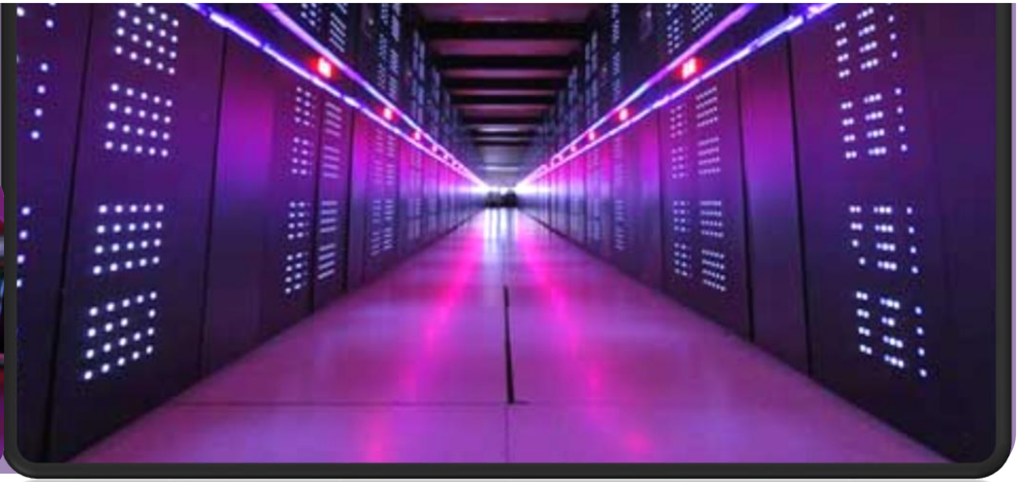
Motivation

Data Movement is the Bottleneck to Performance, Not Flops

Source: Al Geist in “Paving the Roadmap to Exascale”, SciDAC Review 2010

(TH2)

Located in China



Ranked as the world’s fastest supercomputer (Nov. 2015)

- ✓ 33.9 PFLOPS — ✗ has only reached 4% of the exascale target (set for ~2020-2025)
- ✓ 17.6 MW — ✗ has already reached 89% of the 20 MW power limit target *

**P. Kogge. The tops in flops. IEEE Spectrum, 48(2):48–54, 2011.*

Motivation

Data Movement is the Bottleneck to Performance, Not Flops

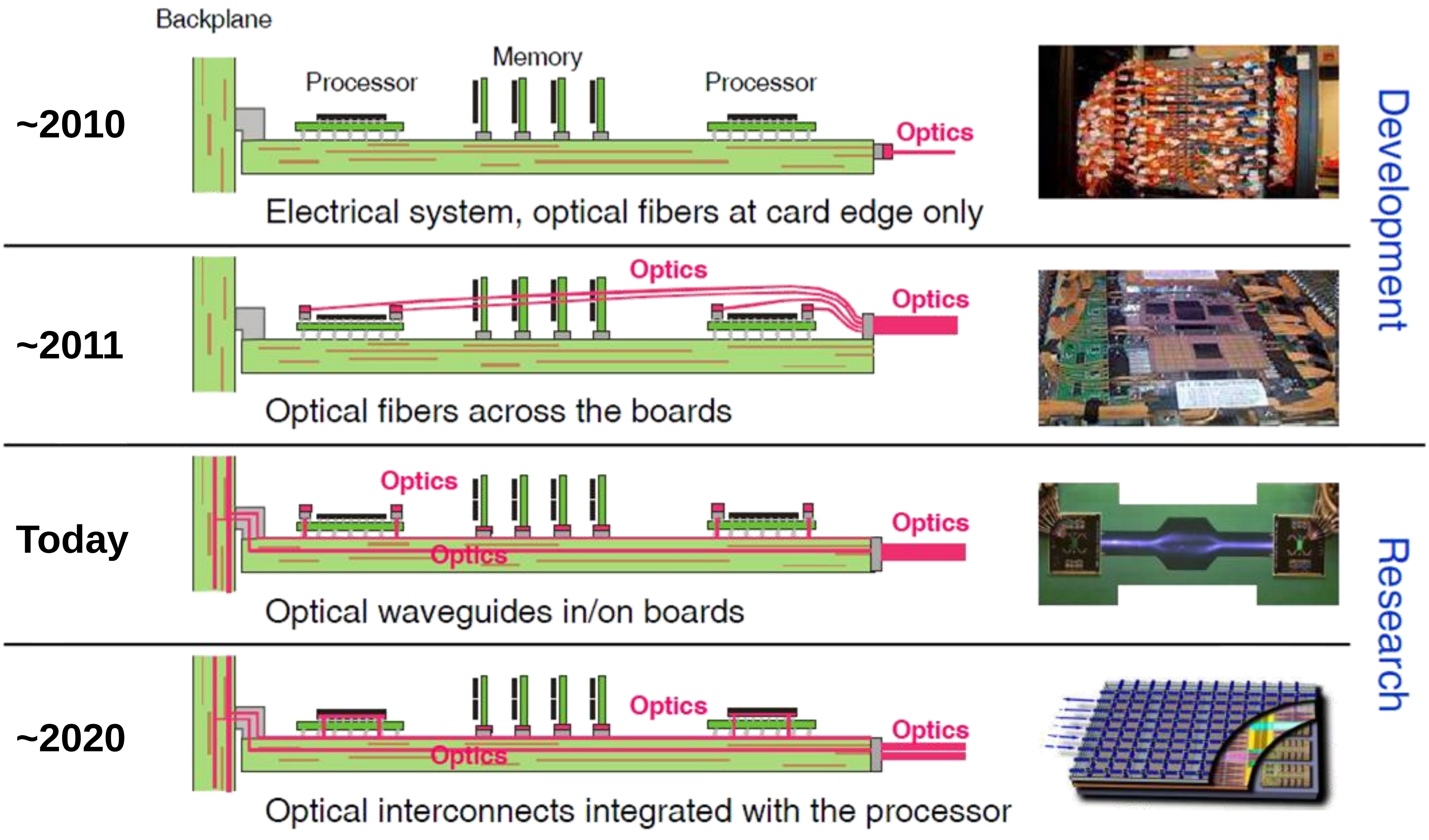
Source: Al Geist in "Paving the Roadmap to Exascale", SciDAC Review 2010

Challenges and the role of Optical interconnects

- ✗ As computation density increases (more cores/chip) leads to higher capacity requirements…
- ✗ …but Copper wires have significant limitations as:
  - they can offer High capacity only for very short distances
  - they present increased power consumption as speed and distance increases
- ✓ Optical interconnects emerge as a promising solution for replacing copper at short distances in future DC and HPC systems
  - they can offer High capacity for both short and higher distances combined with low power consumption

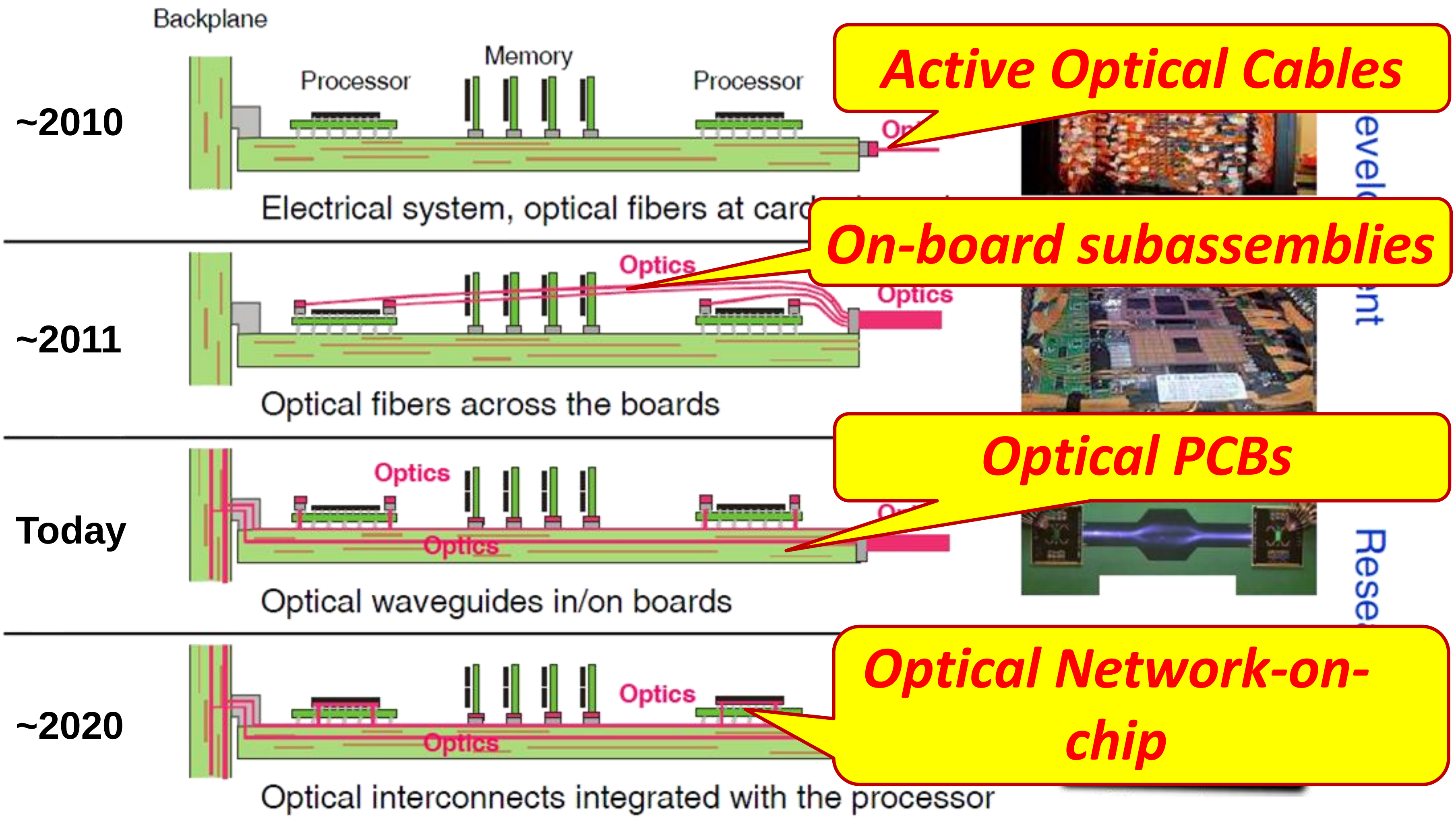
**P. Kogge. The tops in flops. IEEE Spectrum, 48(2):48–54, 2011.*

Optical Interconnects Evolution & RoadMap



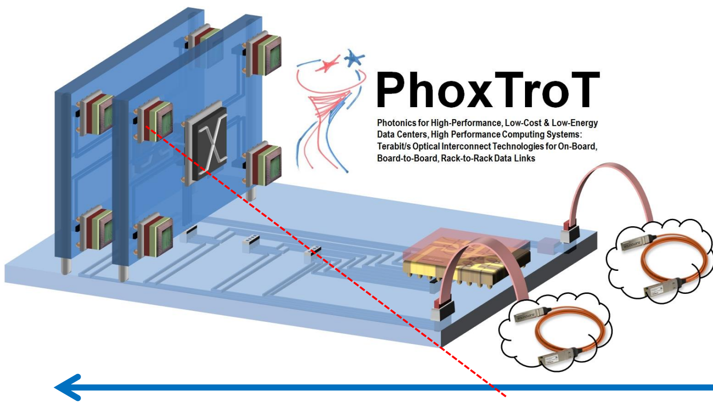
Source: IBM, B. Jan Offrein, "Silicon Photonics Packaging Requirements", Munich 2011

Optical Interconnects Evolution & RoadMap

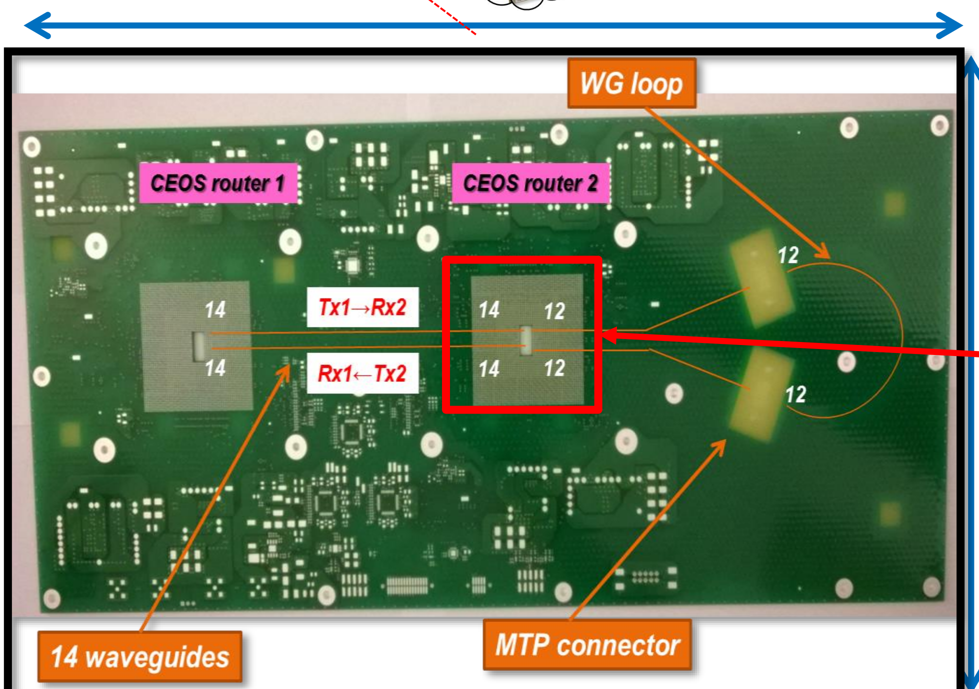


Source: IBM, B. Jan Offrein, "Silicon Photonics Packaging Requirements", Munich 2011

The PhoxTroT Research Project & its Vision



PhoxTroT deals with optical:
(1) On-board,
(2) Board to board and
(3) Rack to Rack interconnects

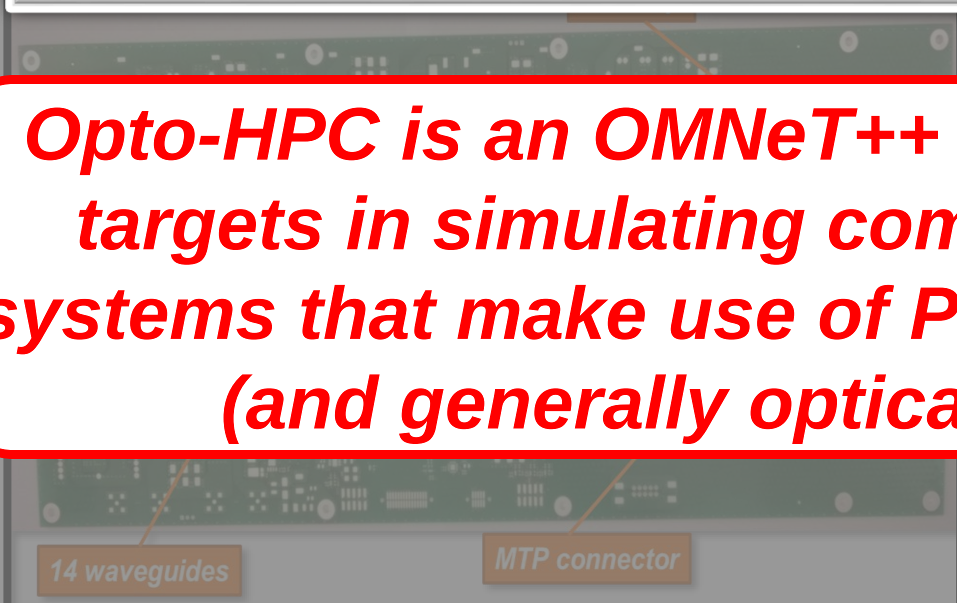


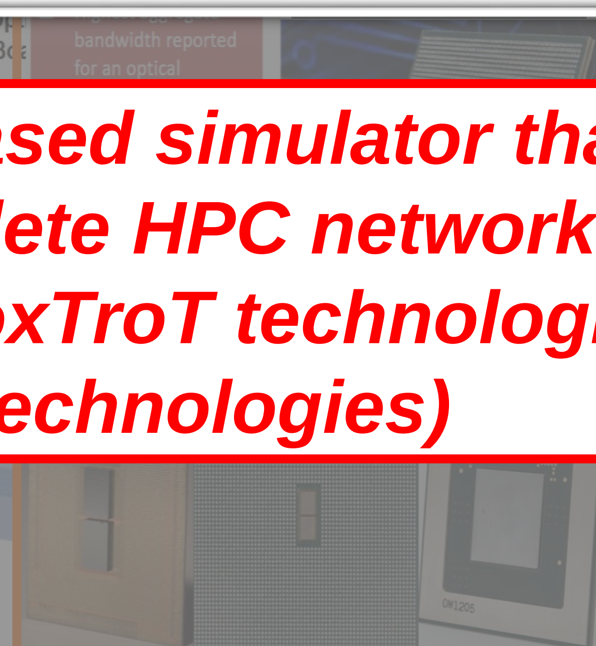
The *OptoHPC* simulator

The PhoxTroT Research Project & its Vision

PhoxTroT deals with optical:

How do all these technology improvements will affect the system-scale performance of an HPC?



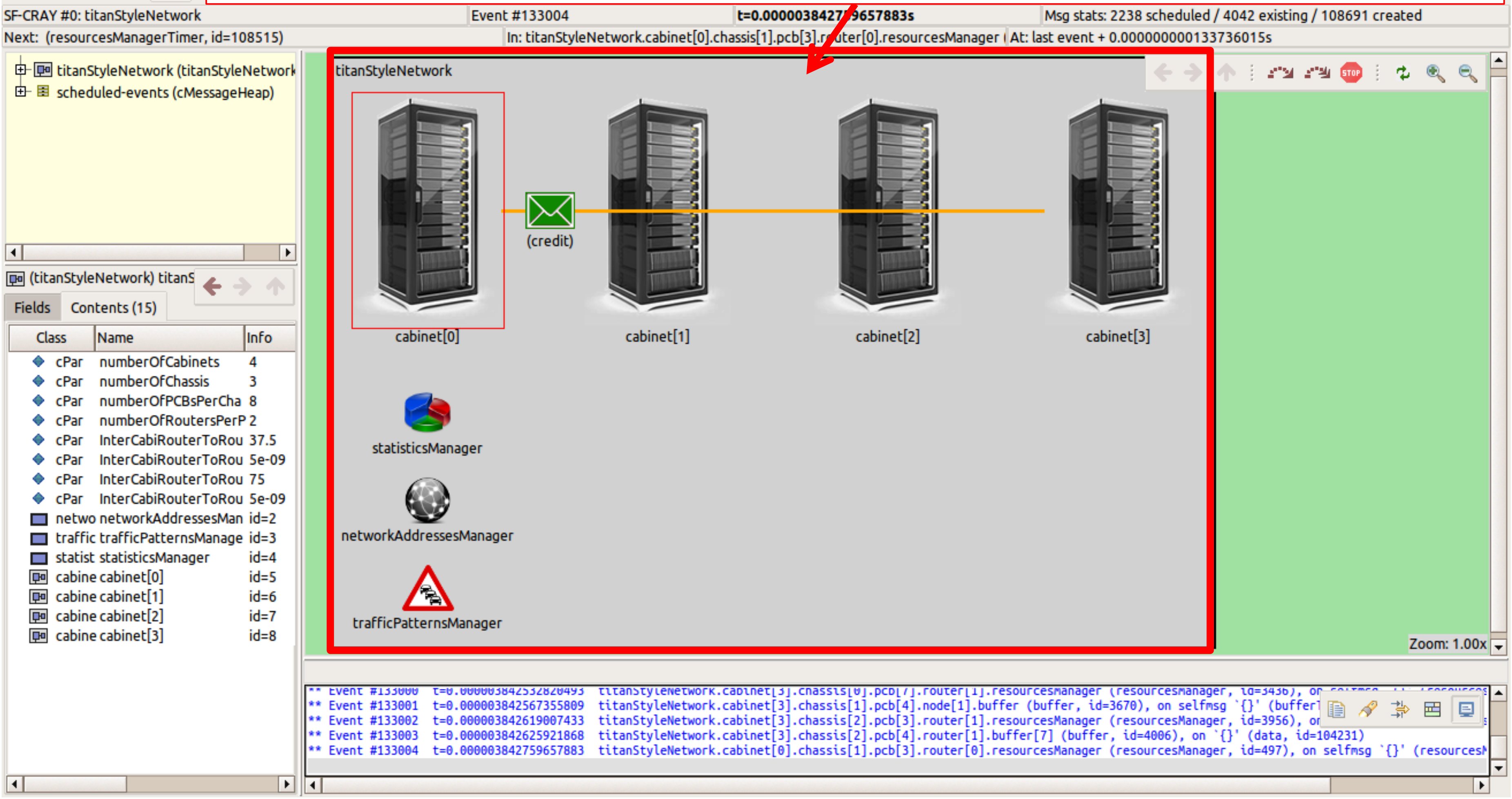


Opto-HPC is an OMNeT++ based simulator that targets in simulating complete HPC network systems that make use of PhoxTroT technologies (and generally optical technologies)

The Op

titanStyleNetwork network module:

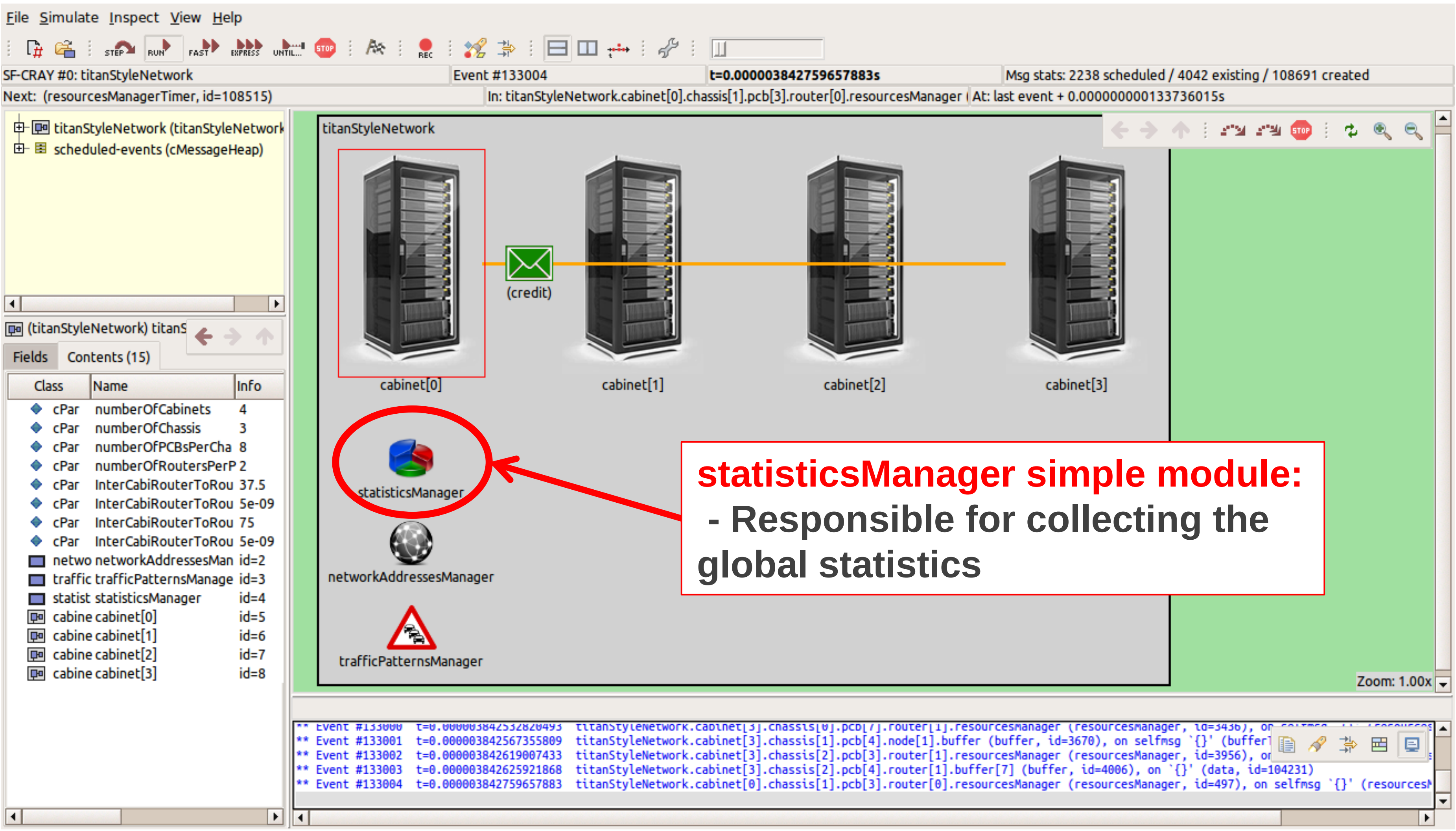
- Defines the connections among the HPC racks and declares the use of the (a) statisticsManager, (b) networkAddressesManager and (c) trafficPatternsManager simple modules
- Can be configured to any **3D Torus** and **Mesh** network desired size



The *OptoHPC* simulator

P/oS-net Research Group

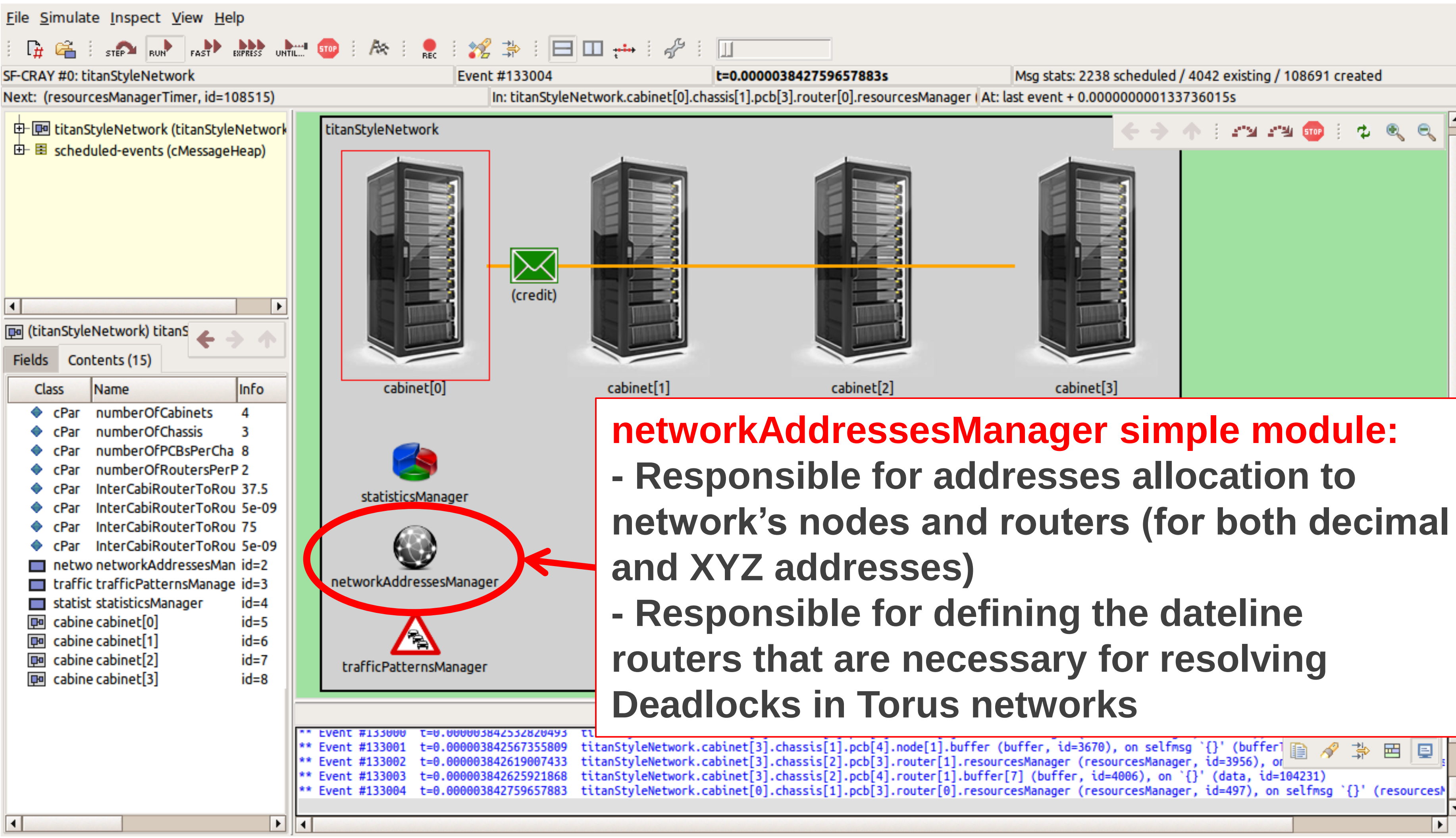
The Opto-HPC simulator



statisticsManager simple module:

- Responsible for collecting the global statistics

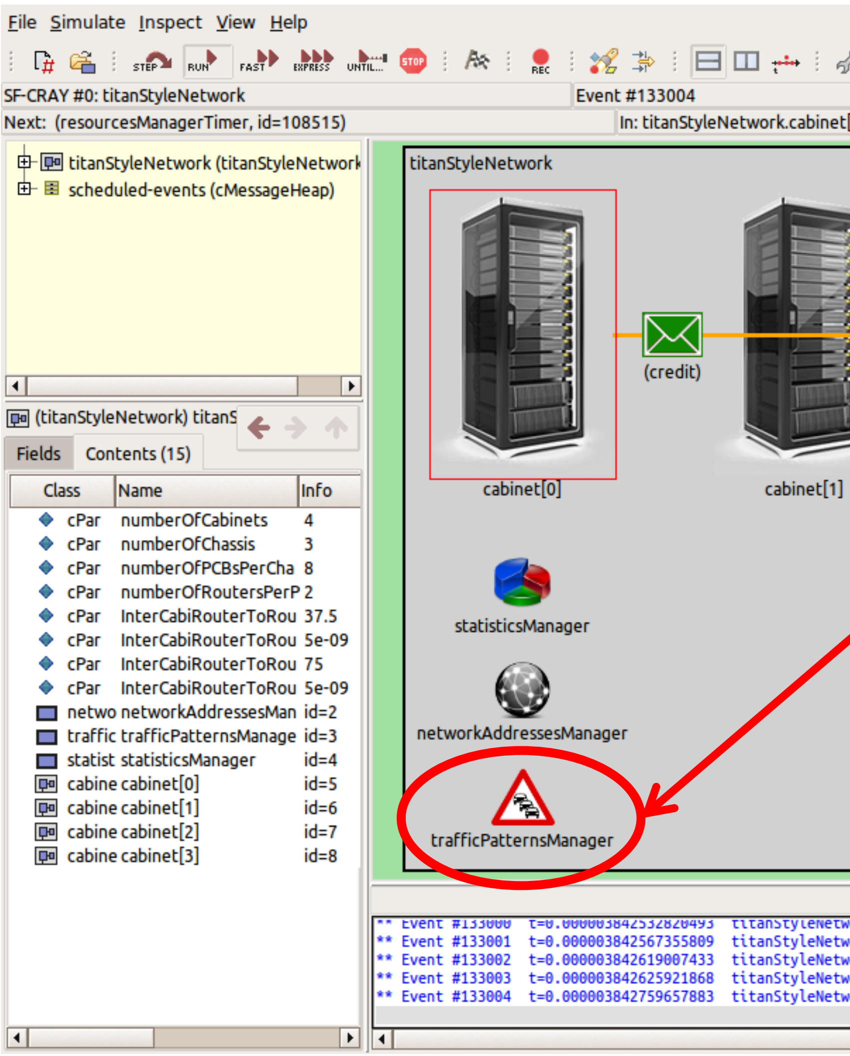
The Opto-HPC simulator



networkAddressesManager simple module:

- Responsible for addresses allocation to network's nodes and routers (for both decimal and XYZ addresses)
- Responsible for defining the dateline routers that are necessary for resolving Deadlocks in Torus networks

The Opto-HPC simulator

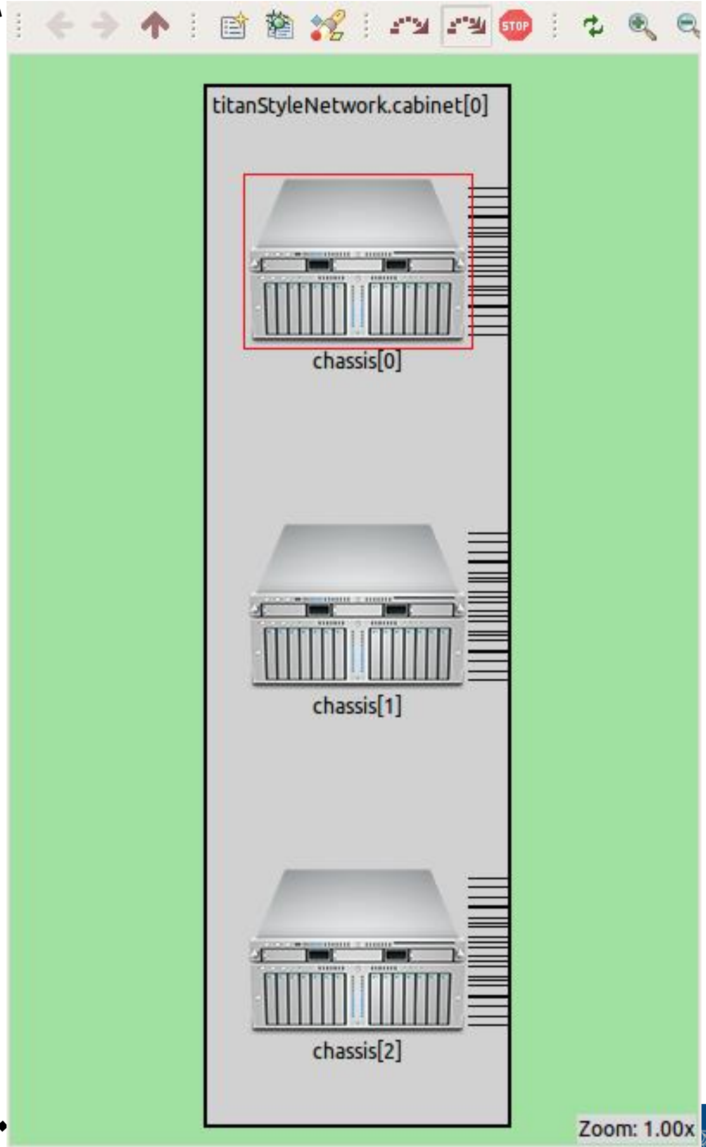


trafficPatternsManager simple module:
Responsible for defining and managing the applications running on the HPC

<u>10 available options:</u>

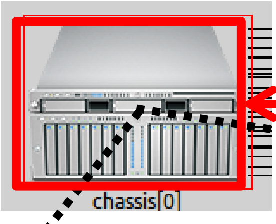
1) Random Uniform
2) Bit Complement
3) Bit Reverse
4) Bit Rotation
5) Shuffle
6) Transpose
7) Tornado
8) Neighbor
9) User defined statistical distributions
10) Packet traces

The Opto-HPC simulator



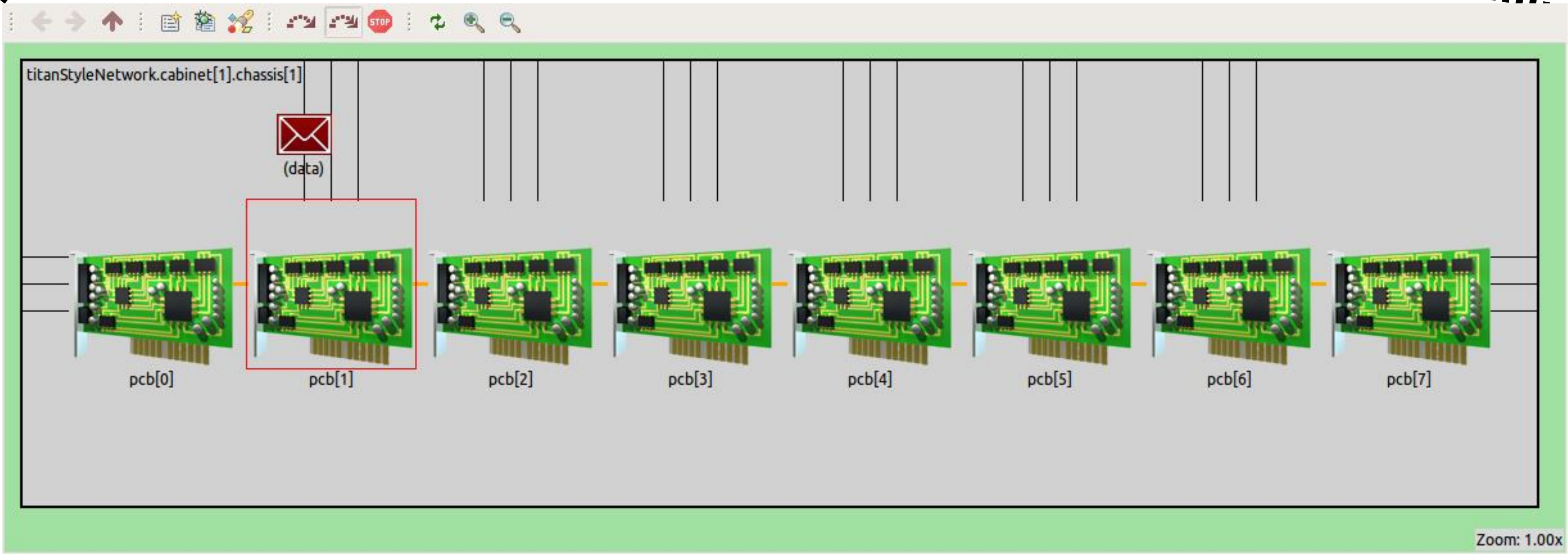
cabinet compound module:
- Defines the connections among the chassis placed in the cabinet and the outer world

The Opto-HPC simulator

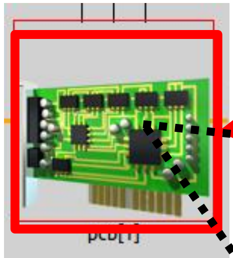


chassis compound module:

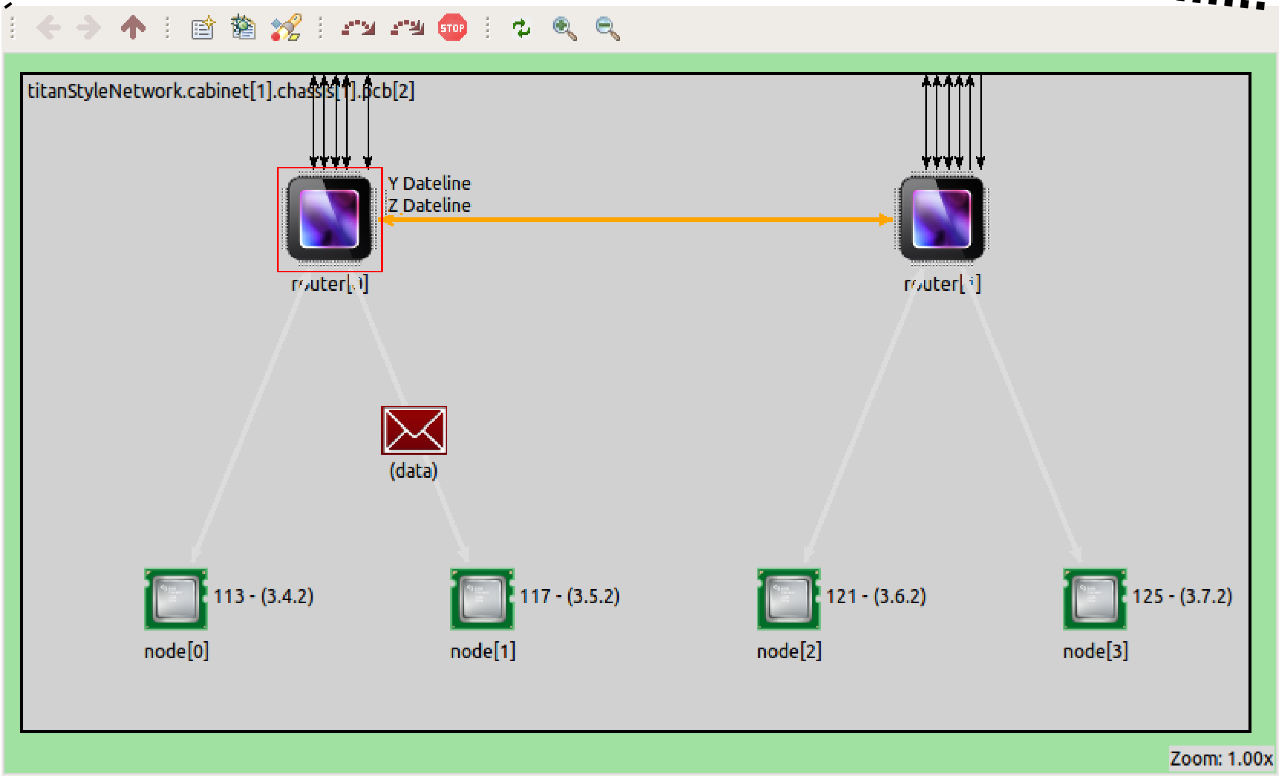
- Defines the connections among the PCBs placed in the cabinet and the outer world



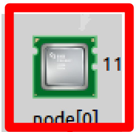
The Opto-HPC simulator



PCB compound module:
- Defines the connections among the nodes and routers inside the PCB and the outer world



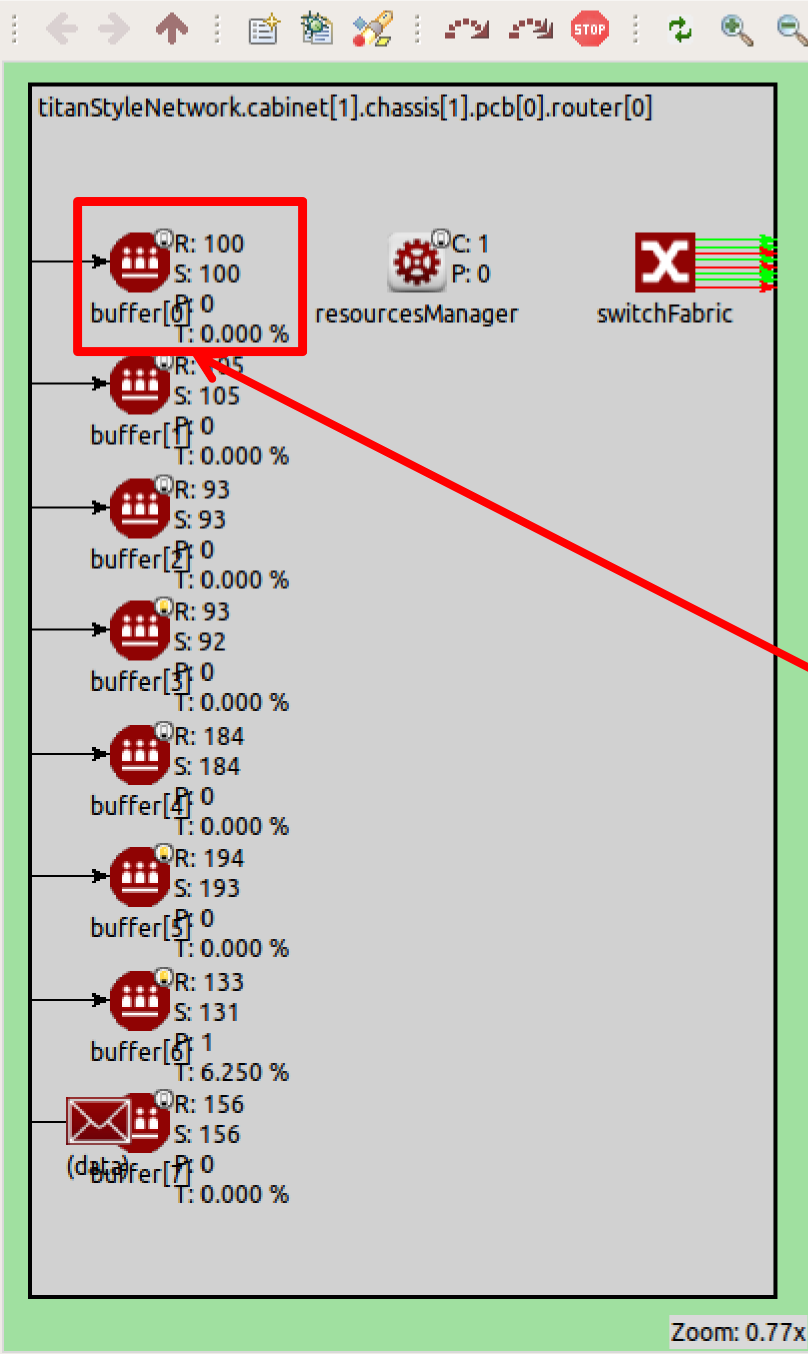
The Opto-HPC simulator

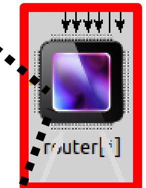


Node compound module:
- Represents the CPU chips used in the HPC
- Embodies all the key simple modules for having "cpu operation"

Router compound module:
- Represents the router chips used in the HPC
- Embodies all the key simple modules for having "router operation"
- Supports **DOR** and **minimal Valiant** routing algorithms
- <u>Utilizes 3 auxiliary classes:</u>

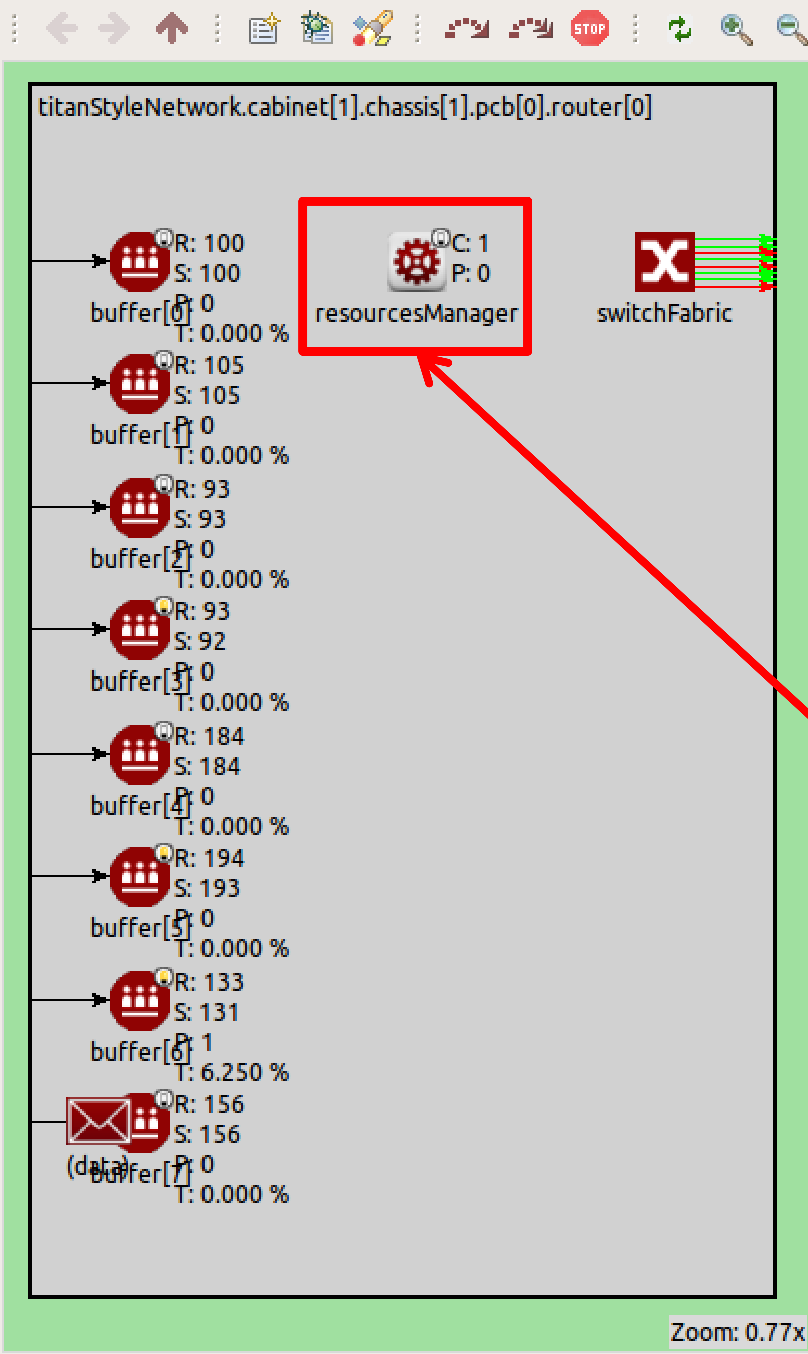
1) shortestPathsManager
2) routingTableManager
3) routingManager

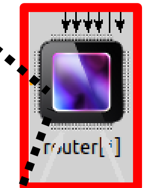




Buffer simple module:

- Implements FIFO queue buffering for the incoming data
- Separated in Virtual Buffers in order to avoid warp-around link deadlocks





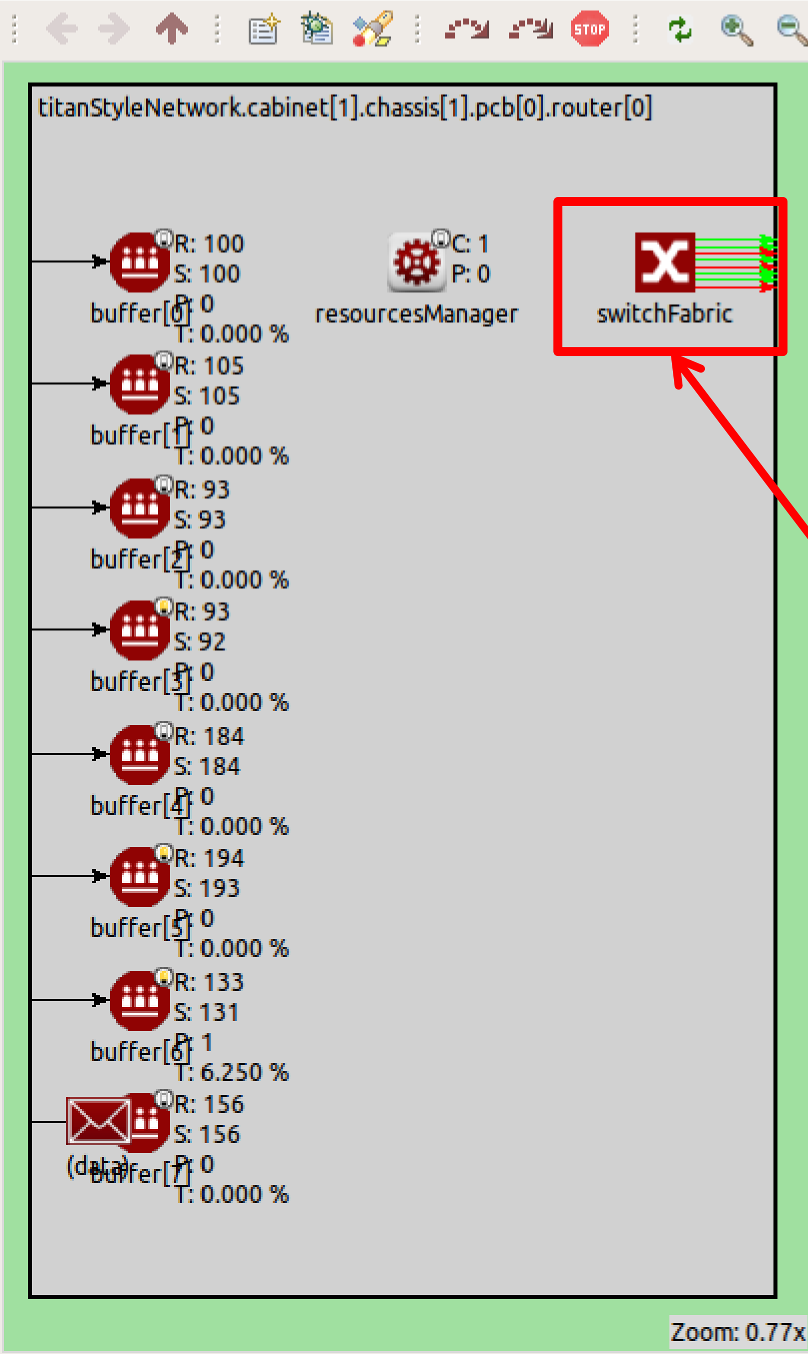
resourcesManager simple module:

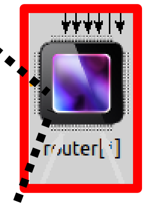
Responsible for:

- the router resources allocation (output ports)
- sending **credit packets** to the previous nodes/routers

Utilizes 3 auxiliary classes:

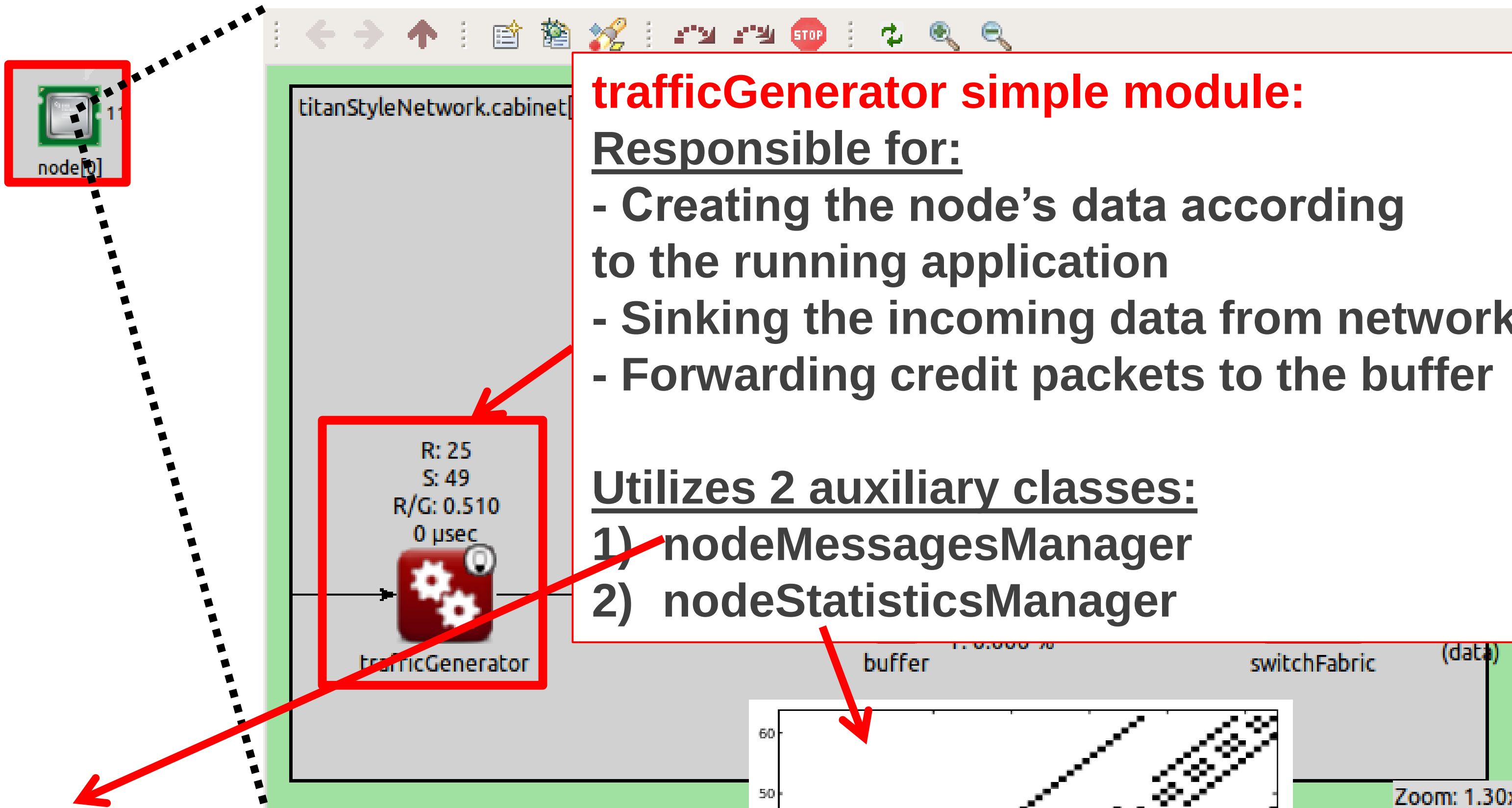
1) pendingDataManager
2) gateAllocationManager
3) creditManager





switchFabric simple module:
Forwards the data transmitted by the buffers/resourcesManager to the proper output port

The Opto-HPC simulator



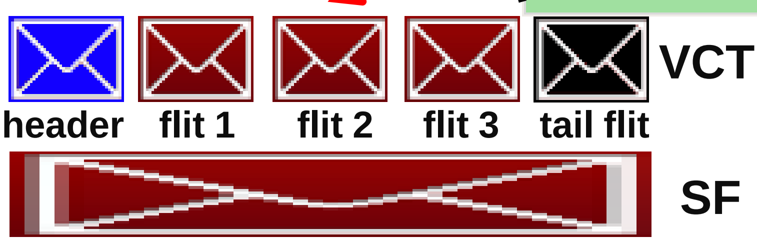
trafficGenerator simple module:

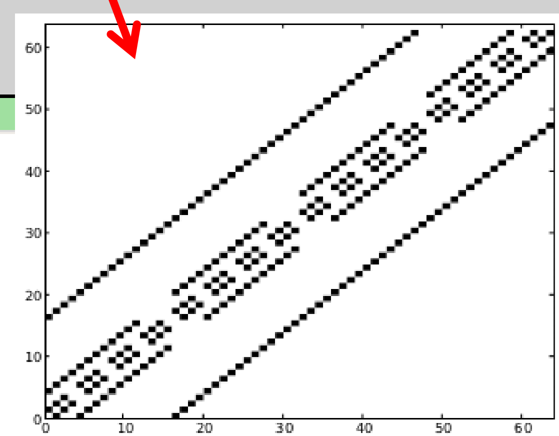
Responsible for:

- **Creating the node's data according to the running application**
- **Sinking the incoming data from network**
- **Forwarding credit packets to the buffer**

Utilizes 2 auxiliary classes:

1) **nodeMessagesManager**
2) **nodeStatisticsManager**





Stats for Nerds

6 Compound Modules
1) titanStyleNetwork.ned
2) cabinet.ned
3) chassis.ned
4) pcb.ned
5) node.ned
6) router.ned

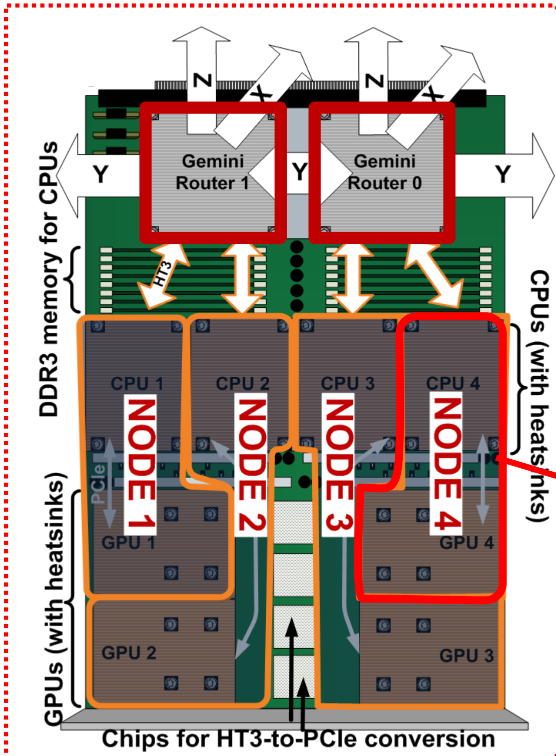
(5 & 6 implement also C++ classes)

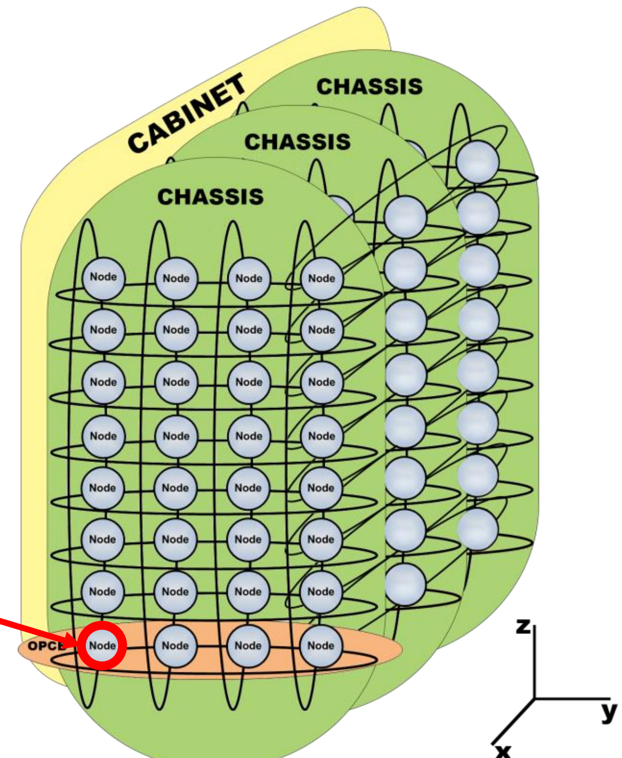
7 Simple Modules
1) networkAddressesManager.ned
2) trafficPatternsManager.ned
3) statisticsManager.ned
4) trafficGenerator.ned
5) buffer.ned
6) resourcesManager.ned
7) switchFabric.ned

5 msg definitions
1) bufferTimer.msg
2) resourcesManagerTimer.msg
3) data.msg
4) flit.msg
5) credit.msg

C++ code
1) 23 new C++ class definitions
2) a total of ~8000 lines of C++ code
3) O(n^2) complexity for the Dijkstra algorithm
4) O(1) complexity for all the major functions (routing decisions, traffic generation etc…)

An *OptoHPC* use case: Titan CRAY XK7 blade vs OPCB





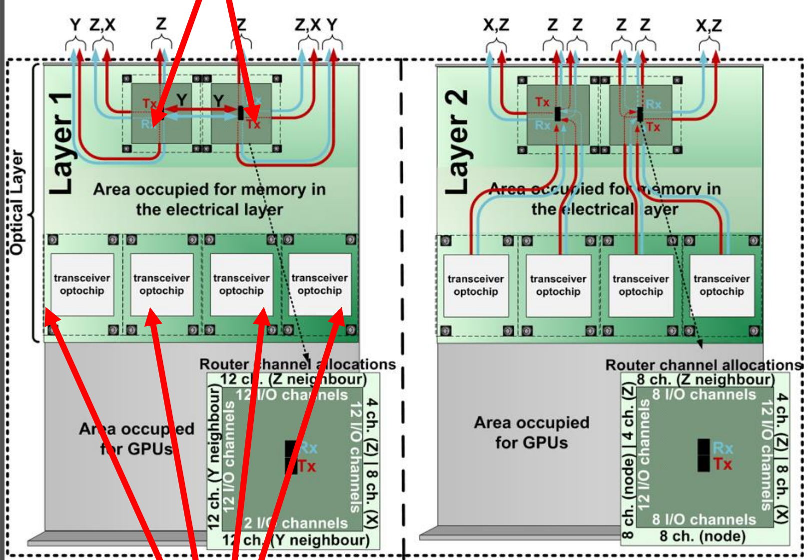
The *OptoHPC* simulator

P/‌oS-net Research Group

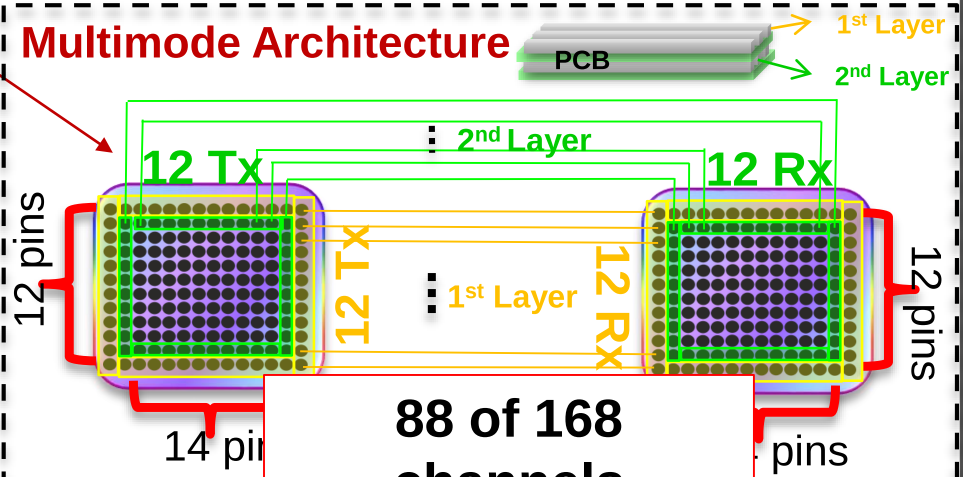
An *OptoHPC* use case: Titan CRAY XK7 blade vs OPCB

O/E routers

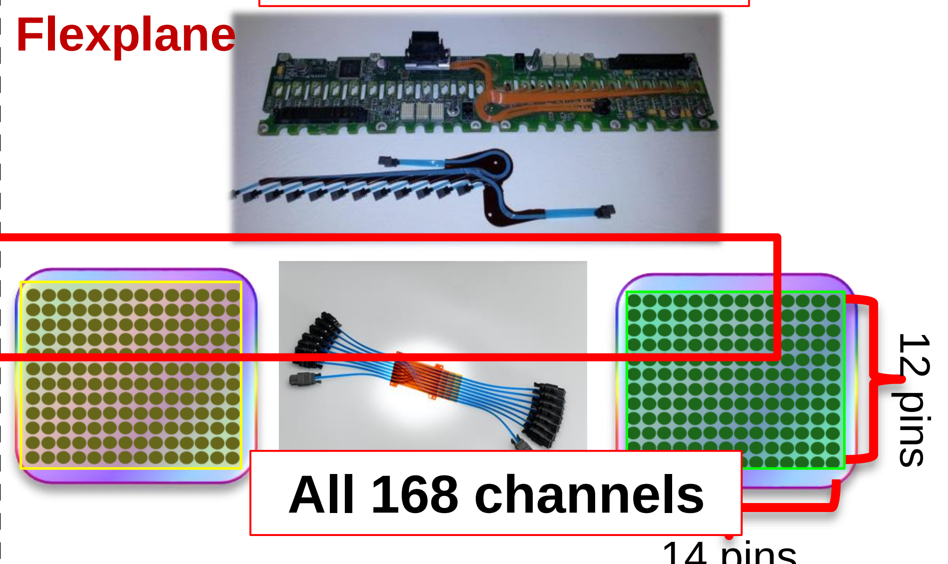
CEOS transceiver matrix



Computing nodes



88 of 168 channels



All 168 channels

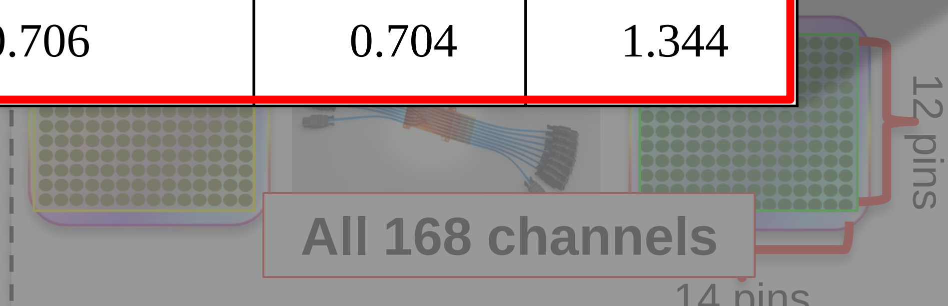
The *OptoHPC* simulator

*Siokis A. et. al. "Laying out Interconnects on Optical Printed Circuit Boards "

An *OptoHPC* use case: Titan CRAY XK7 blade vs OPCB

Router Port Type	Conventional Router	OE-Router-88ch *	OE-Router-168ch *
Node-Router (Gbps)	83.2	64	120
X dimension (Gbps)	75	64	120
Y dimension (Gbps)	75 (Mezzanine) 37.5 (Cable)	96	192
Z dimension (Gbps)	120 (Backplane) 75 (Cable)	128	240
Max Capacity (Tbps)	0.706	0.704	1.344

Computing nodes



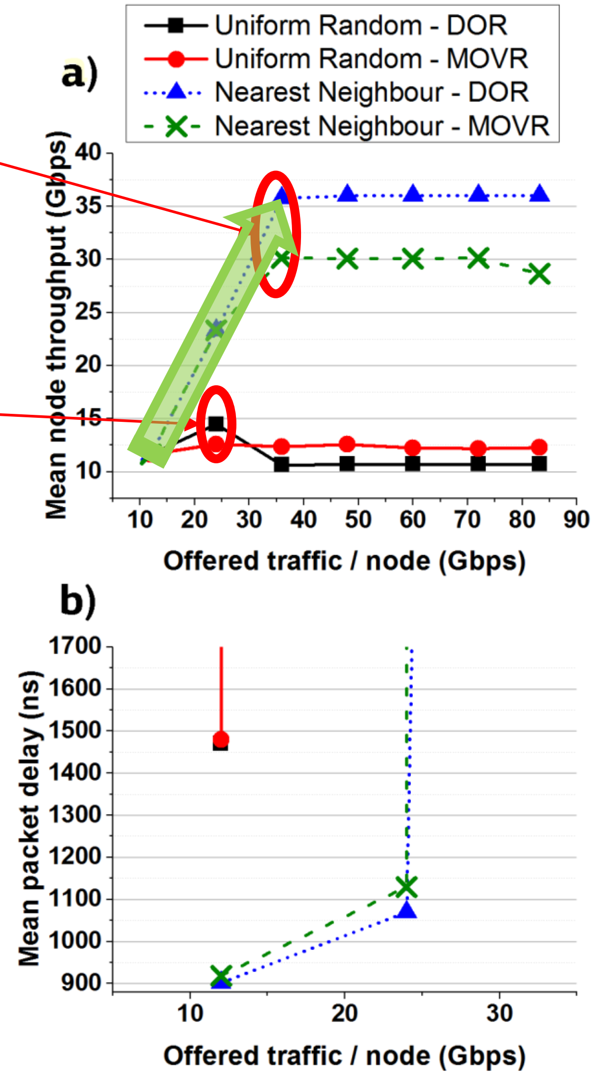
The *OptoHPC* simulator

*Siokis A. et. al. "Laying out Interconnects on Optical Printed Circuit Boards"

Performance Analysis Results – CRAY XK7 for both DOR & MOVR

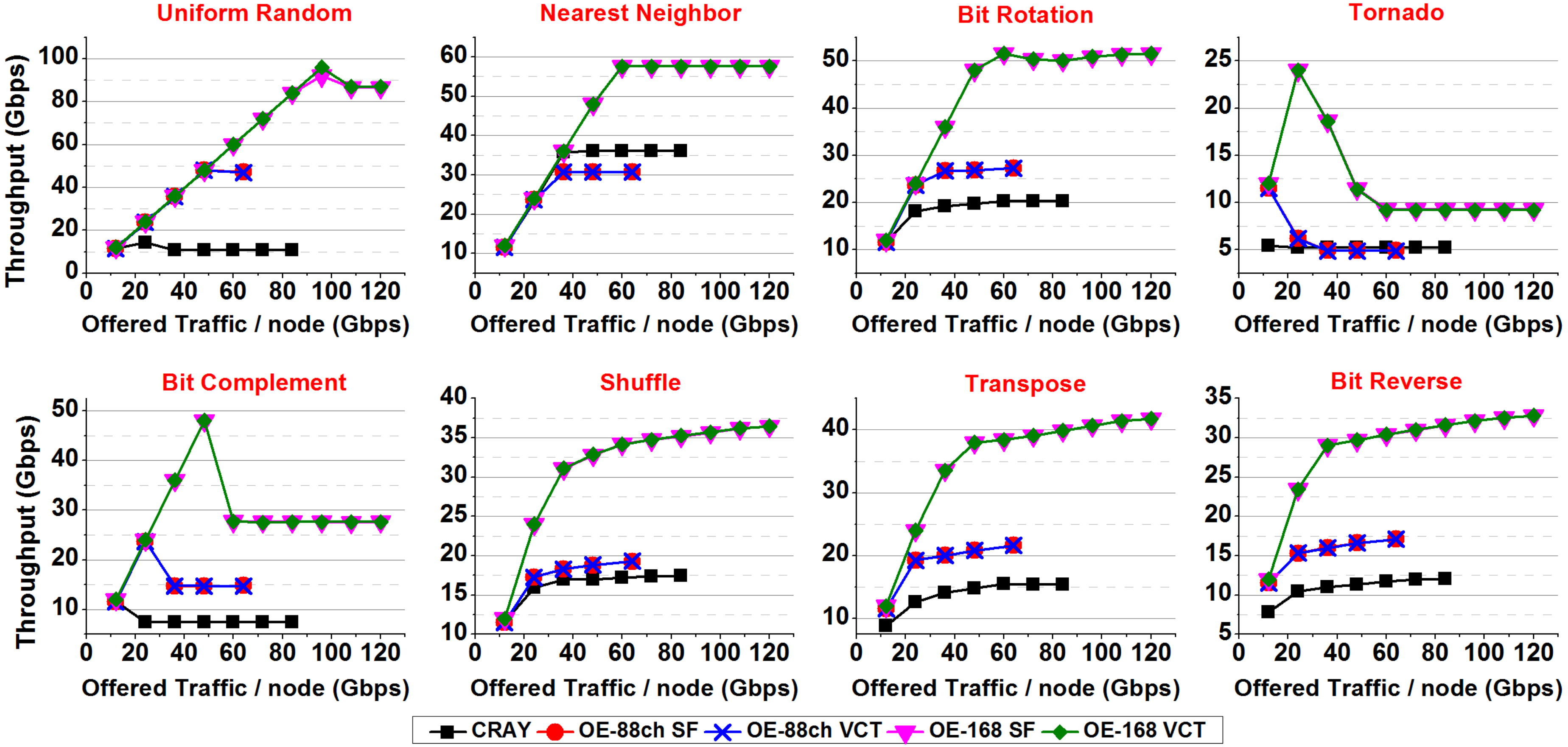
DOR ~20% better

DOR ~15% better



The *OptoHPC* simulator

Performance Analysis Results



The *OptoHPC* simulator

Performance Analysis Results

Mean node Throughput Results			
Pattern	Conventional Router (Gbps)	OE-Router-88ch (Gbps)	OE-Router-168ch (Gbps)
Uniform Random	14.28	48 (3.36x)	92 (6.44x)
Bit Rotation	20.2	27.2 (1.34x)	51.46 (2.54x)
Bit Complement	11.7	23.67 (2.02x)	48 (4.10x)
Bit Reverse	12	17 (1.41x)	32.8 (2.73x)
Shuffle	17.4	19.25 (1.10x)	36.43 (2.09x)
Tornado	5.23	11.51 (2.20x)	24 (4.58x)
Transpose	15.45	21.63 (1.40x)	41.76 (2.70x)
Nearest Neighbour	36	30.7 (0.85x)	57.6 (1.60x)
Mean	~16.5	~24.9 (1.5x)	~48 (2.90x)

Conclusions

Successfully developed a queue-based simulator for complete HPC systems

- ✓ Offers support for both electrical and optical components
- ✓ Currently supports 3D Torus and Mesh Topologies
- ✓ Supports 8 synthetic traffic patterns as well as user-defined statistical distributions and trace files
- ✓ Features both SF and VCT operation like most state-of-the-art routers in the market
- ✓ Implements DOR and Minimal Oblivious Valiant Algorithms (with VC support) allowing for deadlock free operation
- ✓ Comparison between Conventional & O/E technologies using OptoHPC has shown 1.5x mean higher throughput for 88ch. case, 2.9x mean higher throughput for 168ch. case

Thank you for your attention!