

# ATOM Server Configuration

rev2.05

10 April 2014

Satoshi Matsushita



# Photograph of ATOM Chassis

Standard 19 inch and 2U.

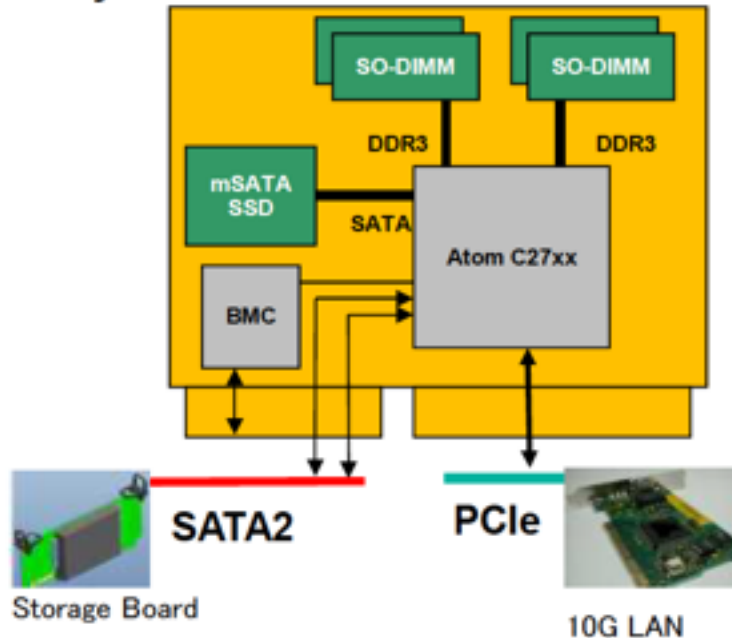
- Each chassis have 46 slots for server blades.
- At 25 degree C, 44 ATOM blades are maximum
- At 40 degree C, 41 ATOM blades are maximum



# ATOM Server Blade

## Server Module

### Block Layout



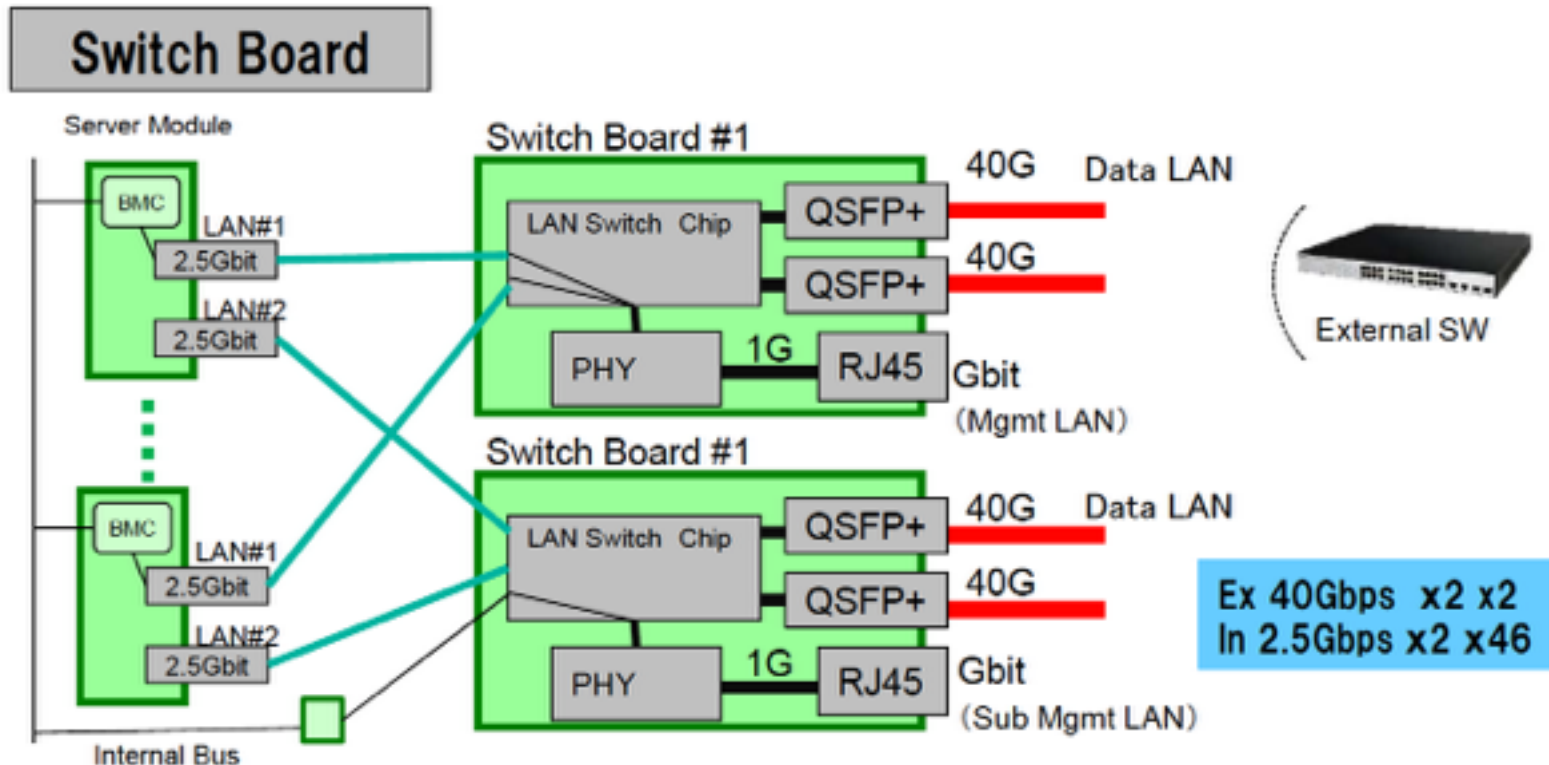
### 【SPECIFICATIONS】

- 1x CPU(Atom™ C27xx)
- 4x SO-DIMM (Max 32GB)
- 1x mSATA SSD (128GB)
- 1x BMC
- 1x SATA3 (To mSATA SSD)
- 2x SATA2 (To storage board)
- 2x 2.5Gbit LAN

Processor	Cores	Frequency	Power
C2750	8C / 8T	2.4GHz	20W
C2730	8C / 8T	1.7GHz	12W

# Connection in a Chassis

1. Chassis come with:
  1. 4 x 3m 40G-40G fibers with server side QSFP+  
40GBase-SR4 multi-mode optical fiber using MPO connector.
  2. 4 x 3m 40G-10G split fibers with server side QSFP+  
Multimode optical fiber.
  3. two IEC C-13 cables for 200V AC power.
  4. slide rails for mount



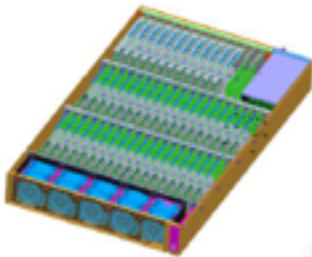
Information herein is preliminary and subject to change.



# Machine Specification

- Three Chassis are arriving on mid April 2014.
- Total: 6U, 150kg
- 200V 6kW power (2 redundant C-13 socket @chassis)
- 4.8kW maximum at normal operation
- less than 1/3 with CPU idle
- Can turn off each server blade and observe chassis temperature with IPMItool.

## Server Chassis



Chassis	
Chassis Dimension (Width x Depth x Height)	2U Chassis Width : 19Inch Depth: 800mm
Weight	Up to 47 kg (system)
Power	Two AC 200V 1.6kw power supply IEC C-14 connector Power consumption : Up to 1.9kw per chassis
Temperature	10 degree C ~ 40 degree C (Ambient air temp.)

# IEC-C13 Rack Mount Power Rail

- Redundancy with each rail to independent wall socket
- Need to identify plug type to floor socket



# Summary of Server Setup

1. Locate a new rack next to existing three racks with Super micro servers and connected to the existing server through 1G ethernet.
2. The new rack consists of:
  1. 1Gbps x 20 port (minimum) LAN switch :
    1. 6 ports from ATOM server
    2. 10 ports for existing Super Micro system experiment
    3. 1 for host machine (rcmaster)
  2. Power rails: Two independent 200V-3kW C14 power rails
    1. 6 x C13 power cables are prepared.
  3. Slide rails for three chassis...
3. 40G optical fibers are provided with the servers
  1. Only QSFP+s for server side are provided
    - 40GBase-SR4 multi-mode optical fiber using MP0 connector
    - The MP0 has 12 cores but SR4 uses only 8 cores.

# Existing Cluster

Three 19 inch APC Racks: <http://www.apc.com/products/family/index.cfm?id=430>

Upper 13U open

Rack with rcmaster: upper 17U open



Rack with Infiniband  
and 10G switch

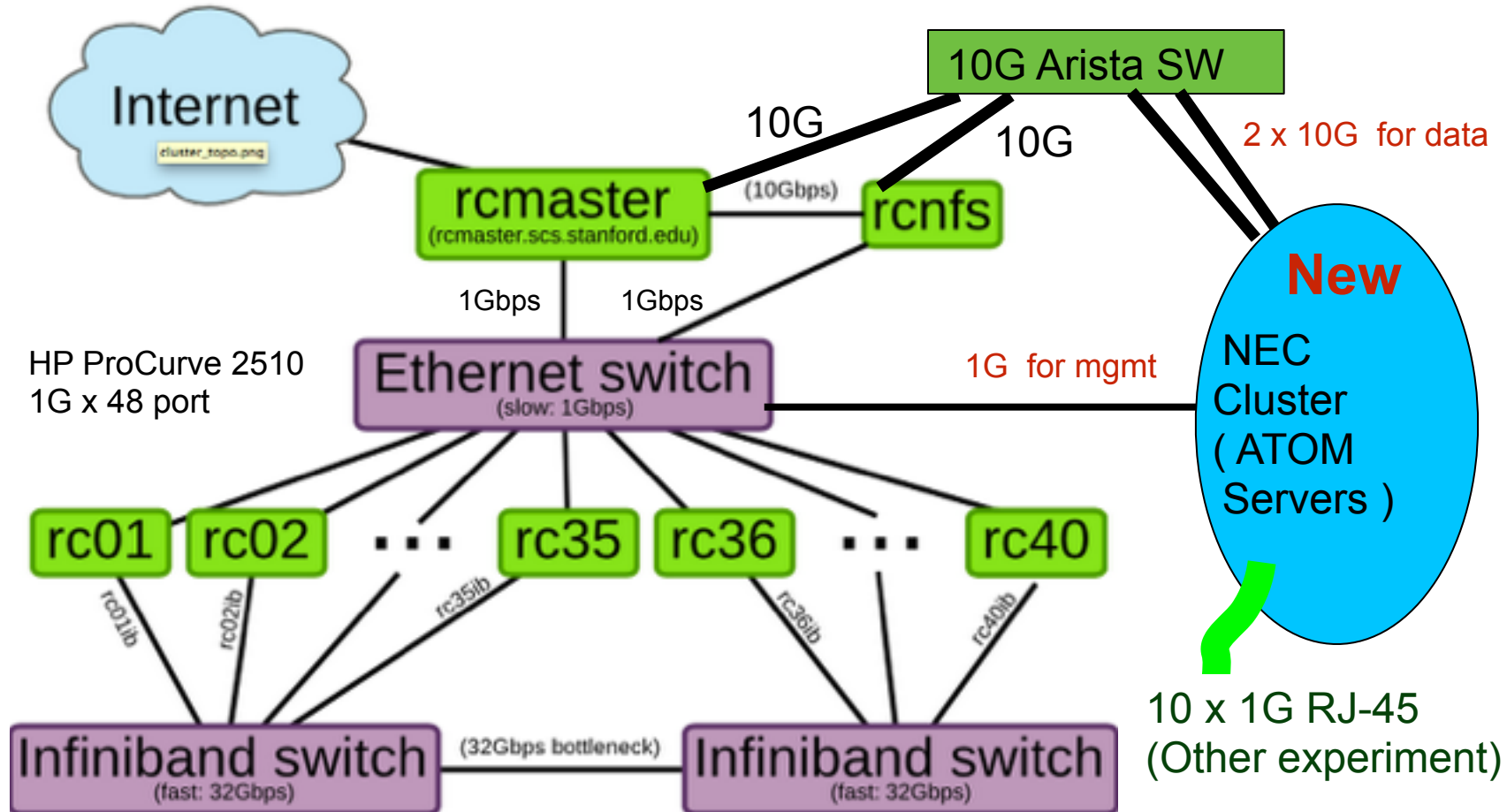


# Network



# Connection in Existing Cluster

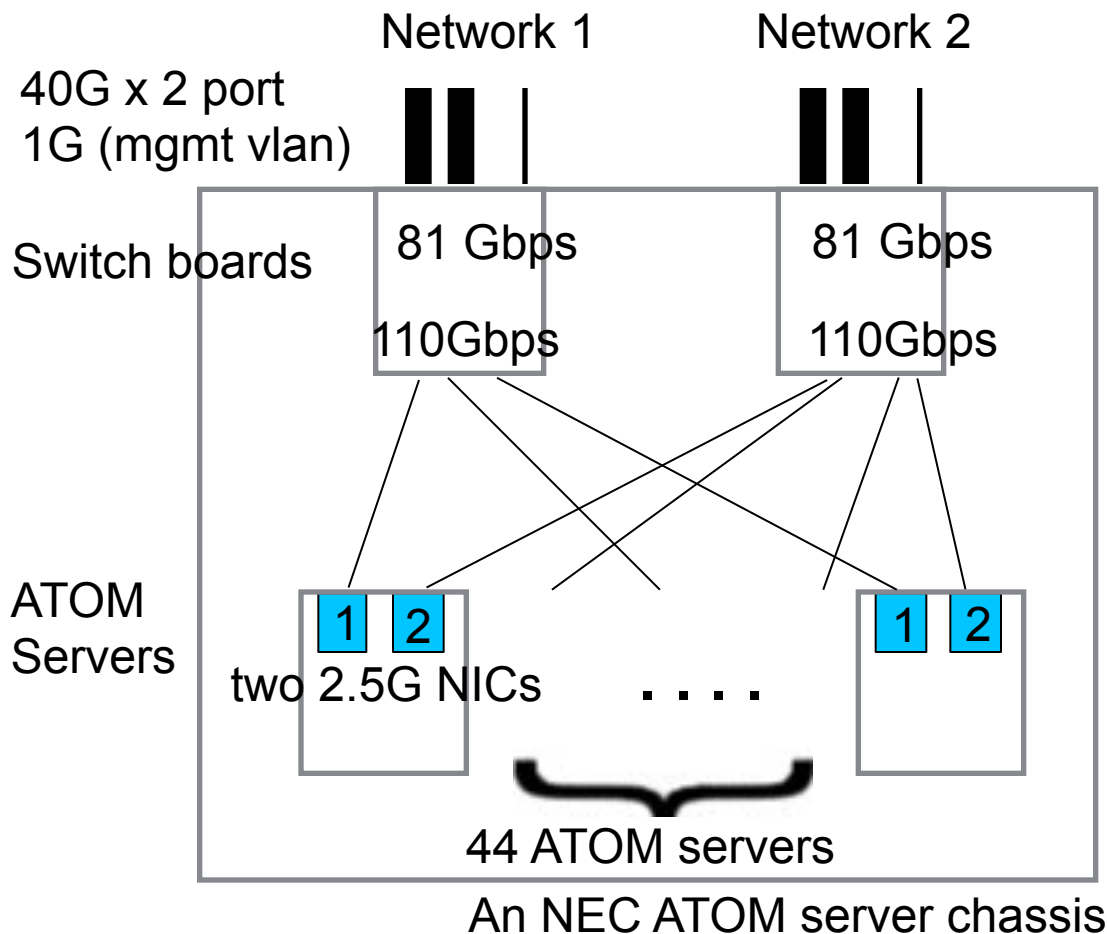
1. rcmaster is host/firewall/DHCP/tftp/IMPI server
2. 10G network exists: a 2 port 10G NIC is installed in host servers.



Cf: <https://ramcloud.stanford.edu/wiki/display/ramcloud/Cluster+Intro>

# Chassis Switches and Two Domains

- Two FM5225 switch boards are installed in a chassis
- Each board consists a separate network domain, i.e.. a packet sent to NIC1 is always delivered to other ATOM server's NIC1 thorough Network1.  
No exchange path between Network 1 and Network 2.



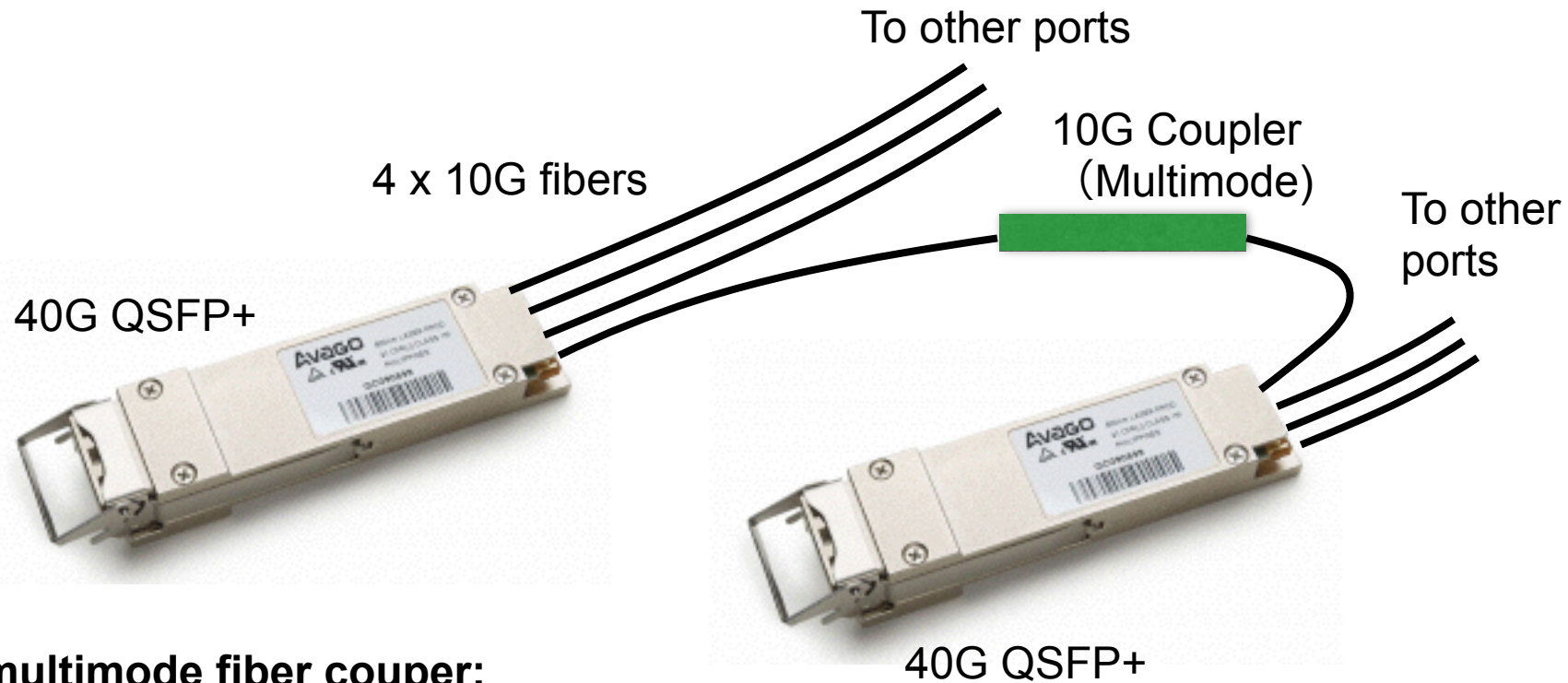
Limitation to create network with chassis switch.

- No spanning tree
- No loop arrowed

(openflow may fix the issue)

# Direct Connection btwn Switch Boards

- Two Fiber sets are provided
  - A. 12 of (40G to 40G fiber) with QSFP+ for chassis side.
  - B. 12 of (40G to 4x10G split fiber) with QSFP+ for chassis side
- Can directly connect switch board in a chassis
  - A. Use 40G QSFP+ for both ends
  - B. Connect 10G with fiber coupler



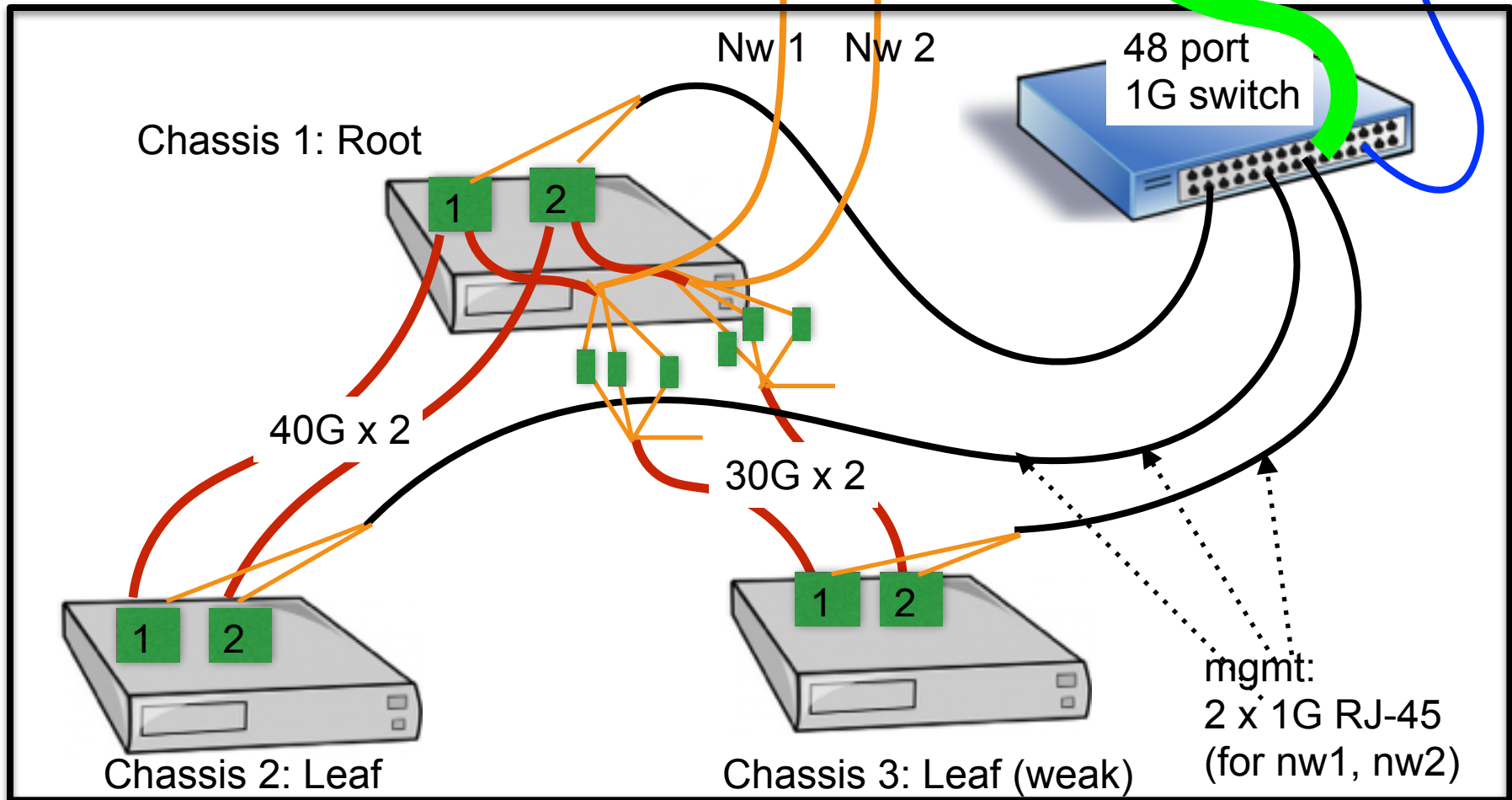
**10G multimode fiber couper:**

<http://www.primuscable.com/store/c/1621-Fiber-Coupler.aspx>

# Connection in NEC Cluster

1. Data path: tree to avoid loop
2. Management: using 48 port x 1G switch

## NEC Cluster



# Future Extension



# Intel® Ethernet Switch FM5224 Microserver Switch Silicon

## Unmatched uServer density

- Up to 72 2.5G ports
- 8 10GbE or 2 40GbE uplinks

## Rapid Array shared memory

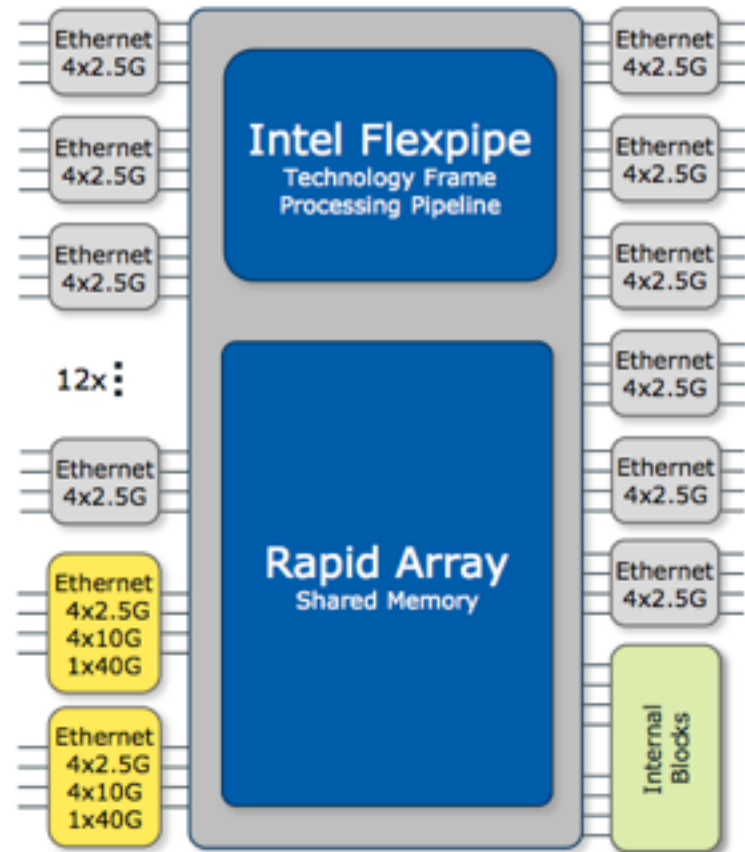
- 8MB shared memory
- 400nS cut-through latency

## Intel® Flexpipe™ Technology frame processing

- Intel Flexpipe Technology frame processing
- VXLAN and NVGRE support
- Advanced load balancing
- IPv4/v6 routing
- CEE/DCB with 8 traffic classes
- Server virtualization support

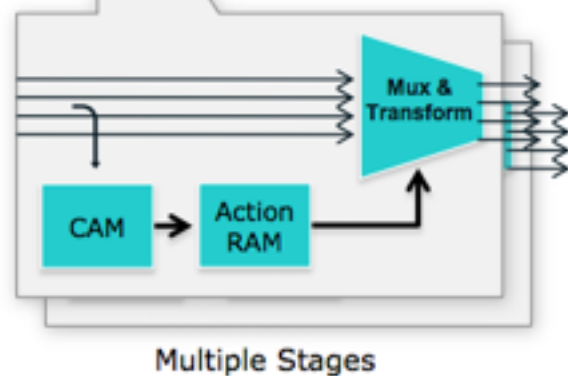
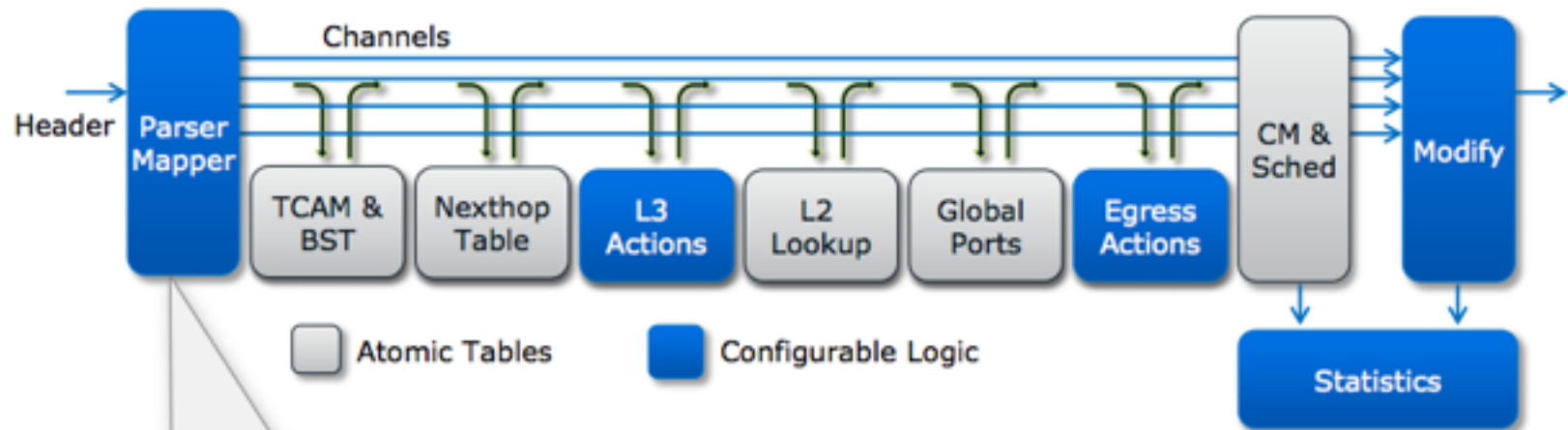
## Compact, flexible port logic

- Integrated SFI, KR PHY
- All ports can also operate at 10/100/1000/2500



IDF13

# Intel® Flexpipe™ Technology Frame Processing Pipeline



## Sample Programmable Protocols

Tunneling	TRILL, MPLS, NAT
Network Overlays	VxLAN, NVGRE
Virtualization	EVB, VEPA, VEPA+, VN-Tag
Proprietary	Customer defined headers

**Programmable and deterministic up to 960Mpps**

**IDF13**



# Configuration without Spine SW - #1

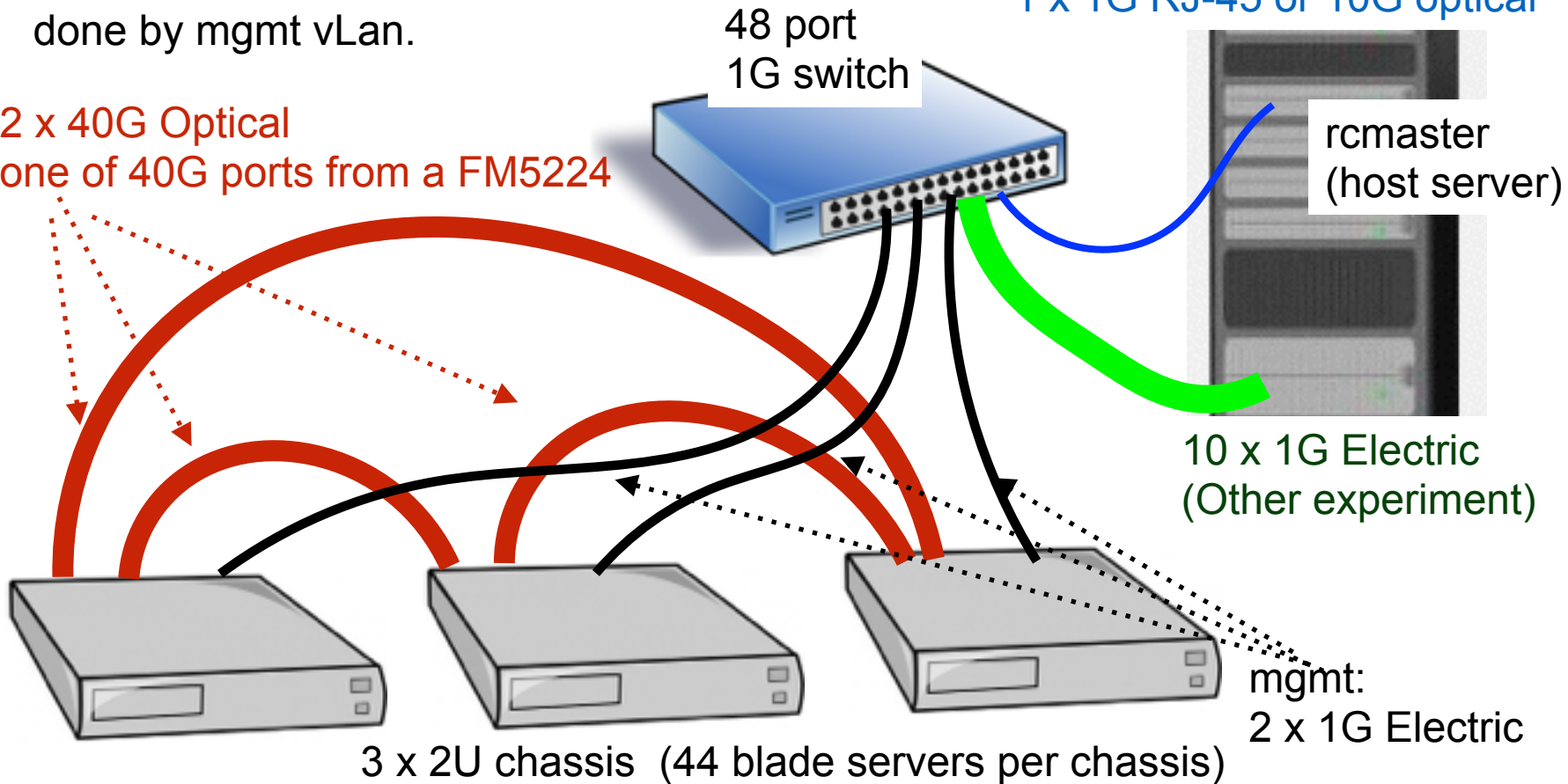
- Intel switch chip FM5224 (TOR in a chassis) is programmable
- Is there any way to connect 40G to host with 10G for performance?

: eg. Using 10G switch with 40G-10G split cable..

Otherwise host access is done by mgmt vLan.

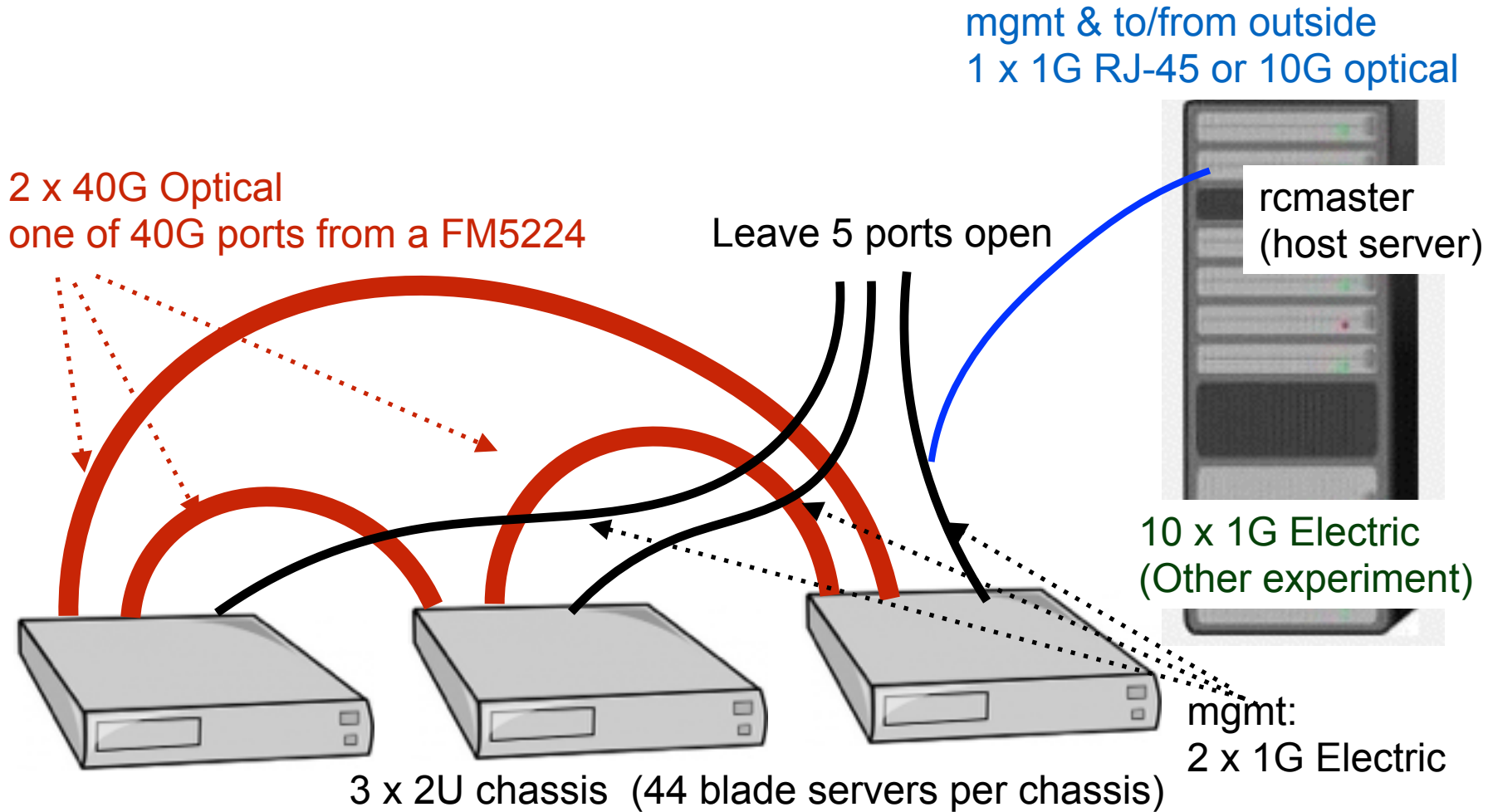
mgmt & to/from outside  
1 x 1G RJ-45 or 10G optical

2 x 40G Optical  
one of 40G ports from a FM5224



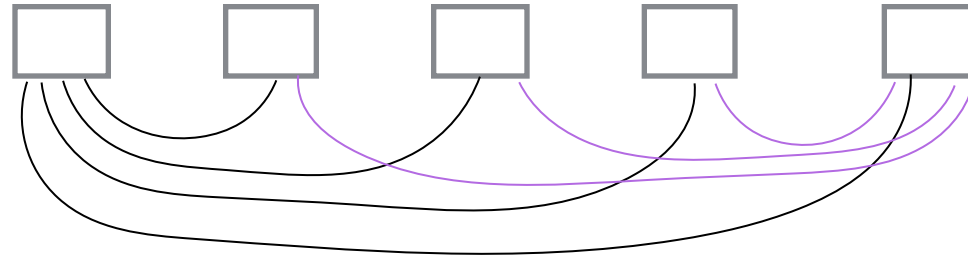
# Configuration without Spine SW - #2

- Program FM5224 to forward management vLan in 40G LAN.
- Does this work?

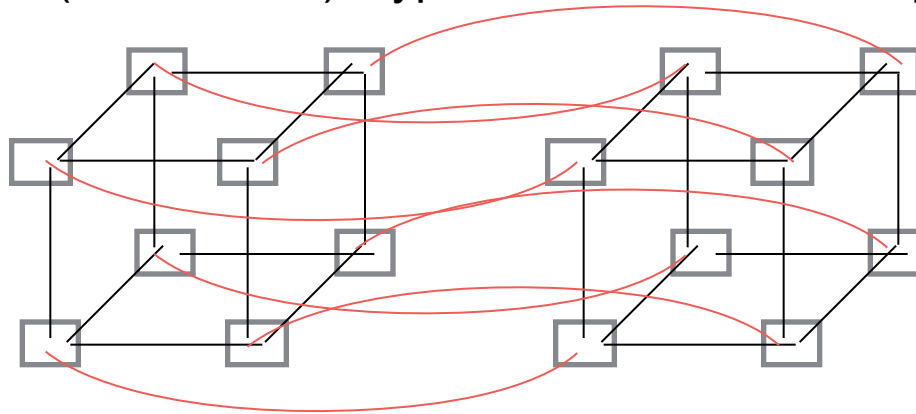


# More Chassis - Various Topology

- Up to 5 chassis (220 servers): full connection with shortest 2 hops - connect other four servers with four links in a chassis



- Up to 16 chassis (704 servers): hyper cube with max 5 hops



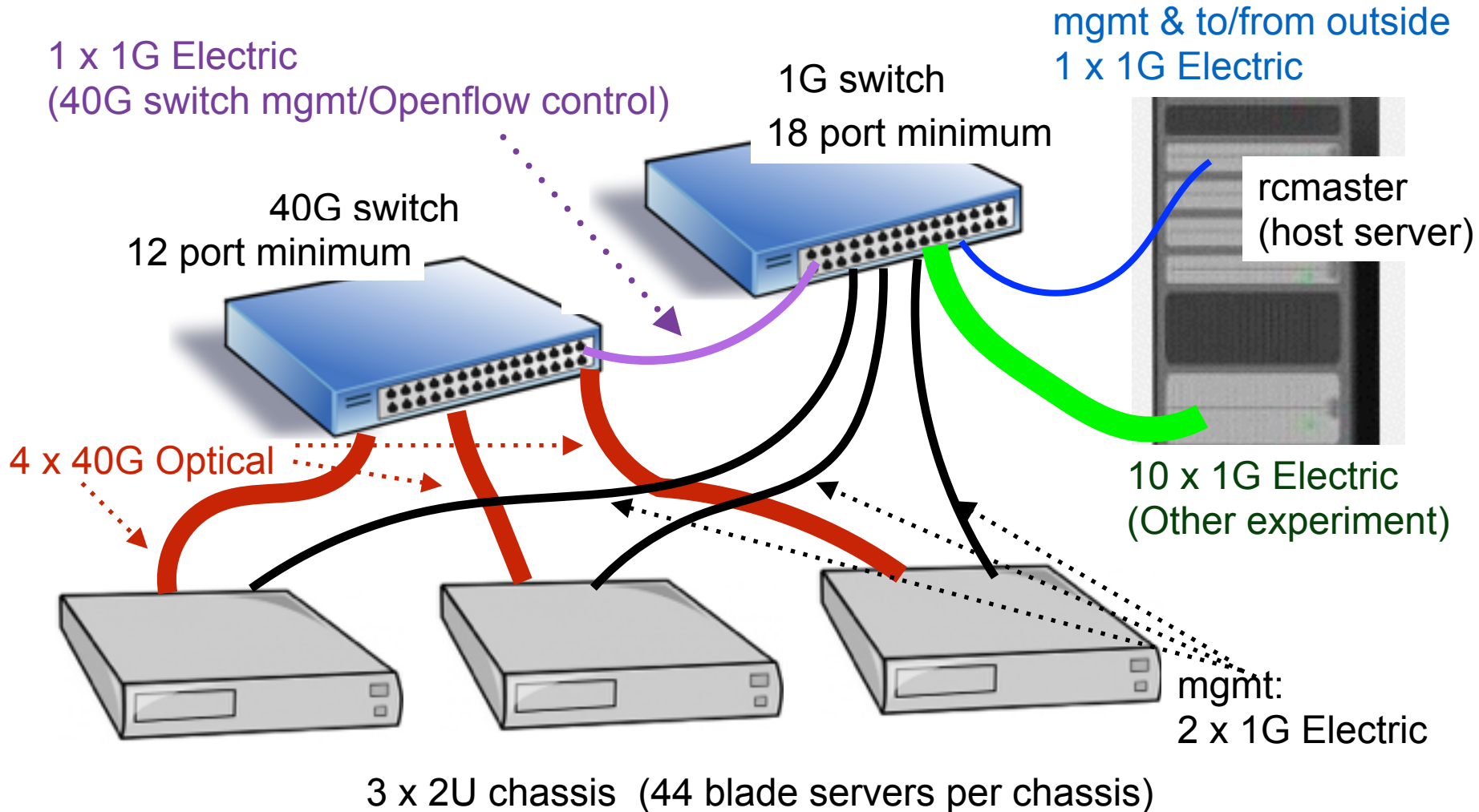
- 4D torus, etc : more chassis with more hops.
- Combination with spine switches

# Previous Plan



# Original Network Connection Plan

- Use rcmaster as host server for DHCP, Firewall, IPMItool, tftp service
- Future experiment with Openflow



# 40G switch candidates

A) For 40G switch, the candidates are Arista 7050X and IBM G8264:

G8264: Low latency, Openflow ready switch.

Cons:

1. 64 port 10G switch (can be used with 40G-10G cables)
2. Support does not seem very well.
3. NEC America carries similar product asking a quotation.

Arista 7050QX-32-F: 32 ports x 40G switch

1. Ultra low latency 550ns per hop
2. High throughput : 1.28Tbps
3. Openflow ready:
  - i. When the flow is on 1500 entry hardware flow table, latency is the same 550ns
  - ii. Z (life time) license needed to start using openflow
  - iii. NICs in ATOM and TOR switch in ATOM server chassis are openflow ready
4. Open software and programmability, base is linux and can be seen as linux server
  1. Additional E (lifetime) license provided for additional functions
5. Good observability
6. Better support. headquarter in Santa Clara
7. Can operate 100V to 250V AC



# 1G switch candidates

1. Use consumer grade switch.

eg. NETGEAR ProSAFE 24-Port Gigabit Ethernet Rackmount Switch (JGS524NA)  
<http://amzn.to/1homYG5>

Amazon Price: \$174.99 (List Price: \$335.00)

2. Use server grade switch. 48x1G/100M switch with 4 x 10G ports.

Web: <http://www.aristanetworks.com/en/products/7048>

Datasheet: <http://www.aristanetworks.com/media/system/pdf/Datasheets/7048T-A%20DataSheet.pdf>

Pros)

- i) 48 x (1G/100M) RJ-45 port, 4x10G SFP/SFP+ (Fiber) ports
- ii) low latency: 3us for 64B frame
- iii) same software visibility as 40G switch.
- iv) server grade, redundant AC (100-250V) and Fans

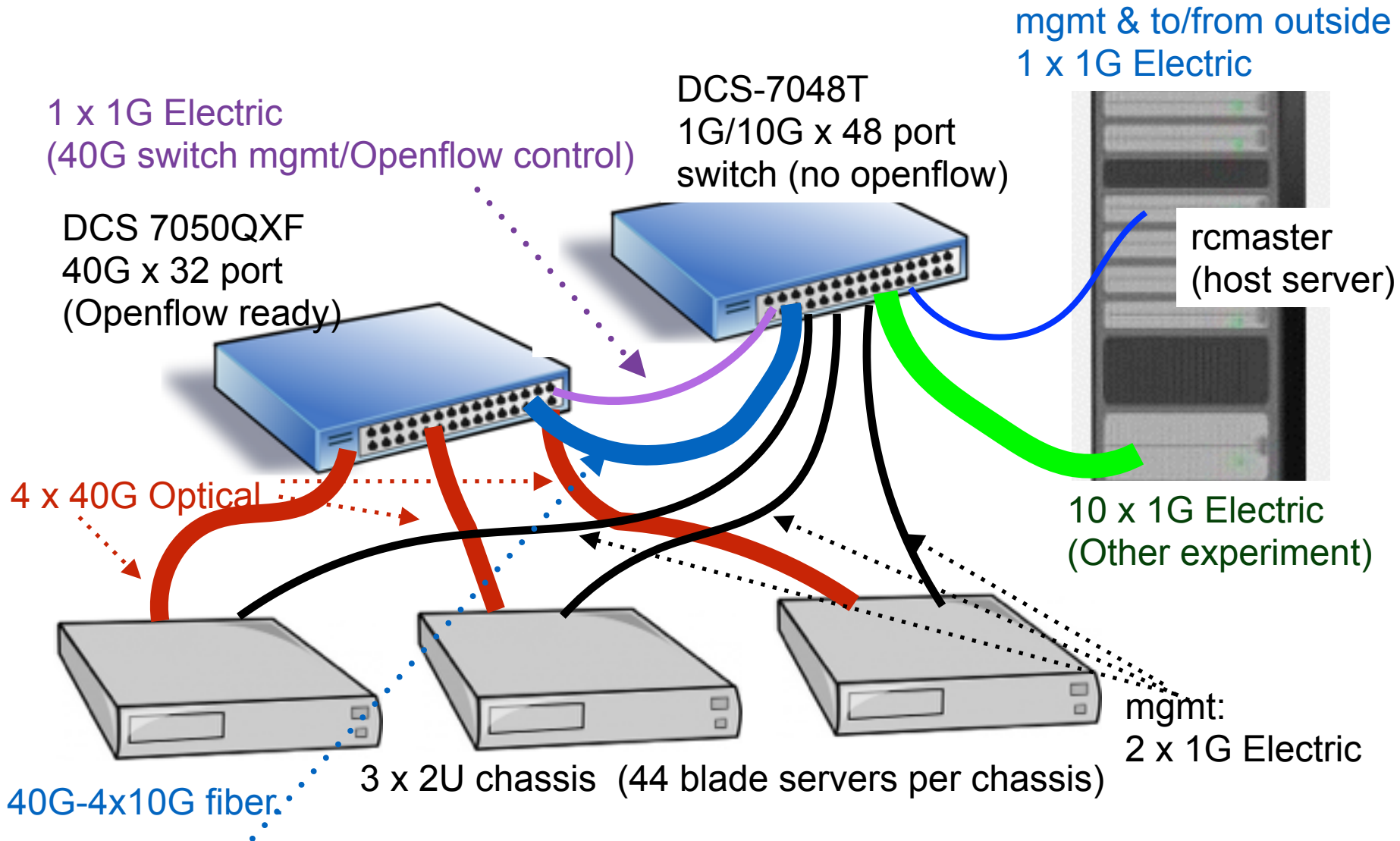
Cons)

- v) not Openflow ready (using rather old generation switch LSI)
- vi) expensive: \$8,400



# Configuration with Arista Switch

- Assume openflow enabled on 40G switch.



Can we directly exchange packet between 40G to host through DCS-7048T?



End

