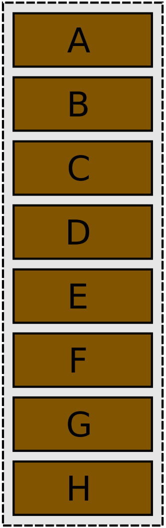
Towards Reconfigurable Rack-Scale Networking

Tyler Szepesi, Bernard Wong, Tim Brecht, Sajjad Rizvi

Cheriton School of Computer Science
University of Waterloo

April 21, 2015

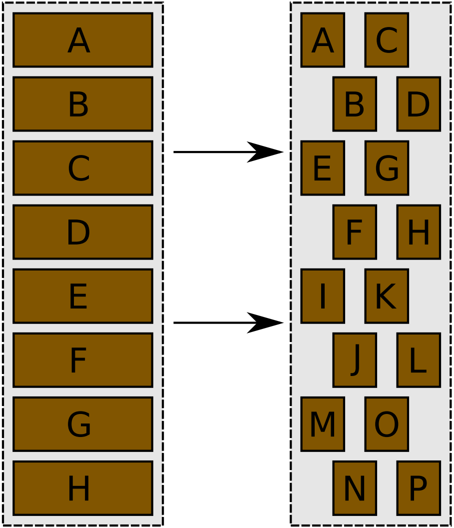
Rack-Scale Computing



Traditional Rack:

- 10s of servers
- 10s of Gbps per server

Rack-Scale Computing



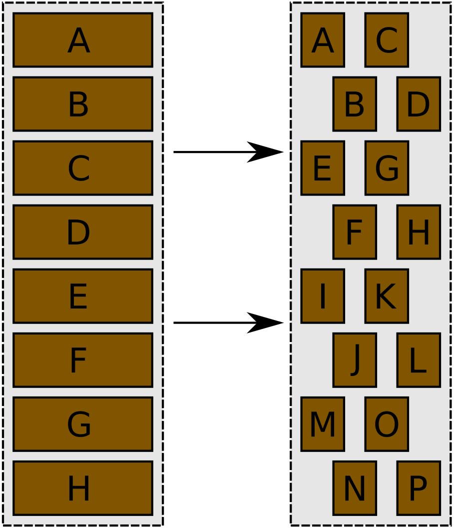
Traditional Rack:

- 10s of servers
- 10s of Gbps per server

Rack-Scale Computing:

- 100s of micro-servers
- 100s of Gbps per micro-server

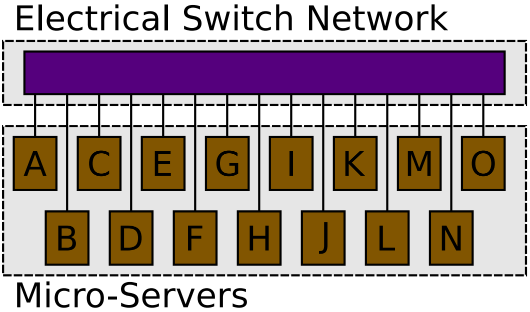
Rack-Scale Networking



A key enabler of rack-scale computing is a network fabric that provides high-bandwidth in a cost effective way.

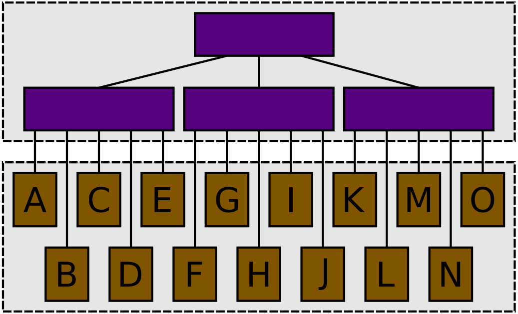
What is the right network fabric?

Single Switch



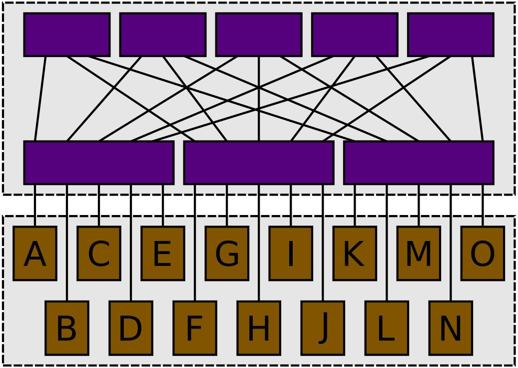
Requires hundreds of ports at hundreds of Gbps per port

Oversubscribed Tree



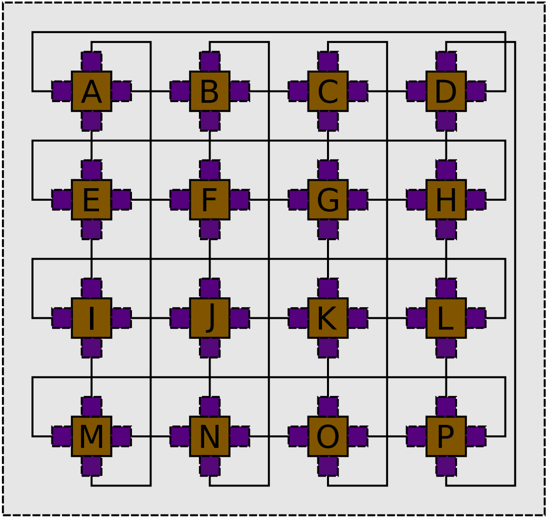
Limited bandwidth for many communication patterns

Fat-tree (Folded Clos)



Costs almost as much for the switching hardware as the micro-servers being networked together

Distributed Switching (Torus Networks)



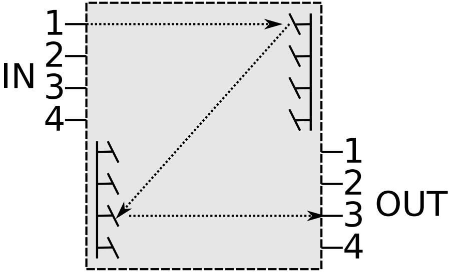
A tradeoff between long path lengths and high port counts per micro-server

Reconfigurable Networks

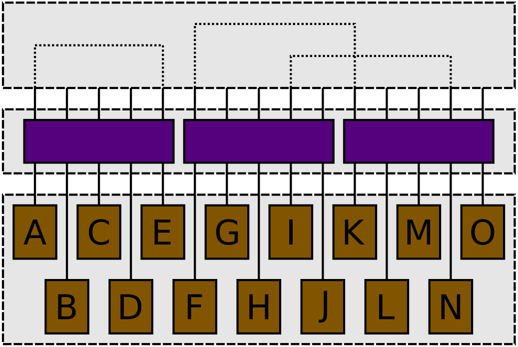
Provide bandwidth where it is needed, when it is needed, and minimize over-provisioning

Optical Circuit Switching

- High bandwidth
- Low cost
- Low power consumption



Optical Circuit Network

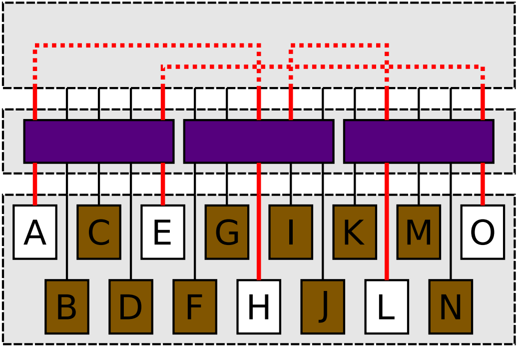


Most effective when the communication pattern between switch changes slowly

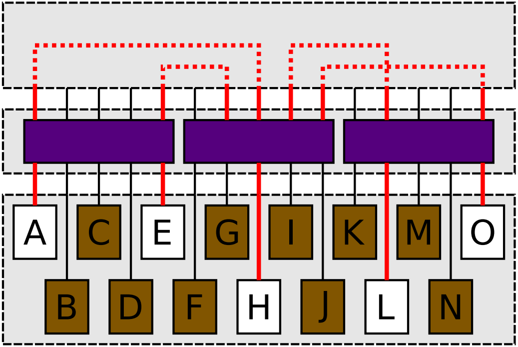
Rack-Scale Communication

- The expected pattern of communication:
  - Groups of micro-servers are used for a task
  - New groups are formed for new tasks
- High bandwidth is needed between members of the group
- Minimal bandwidth is needed for inter-group communication

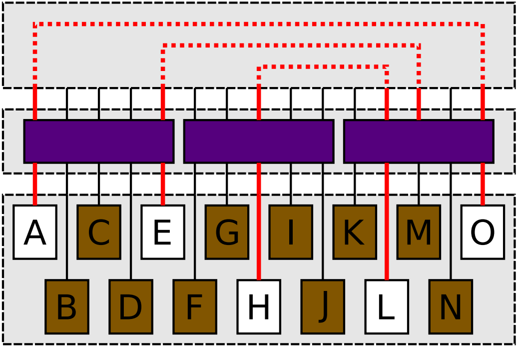
Optical Interconnects



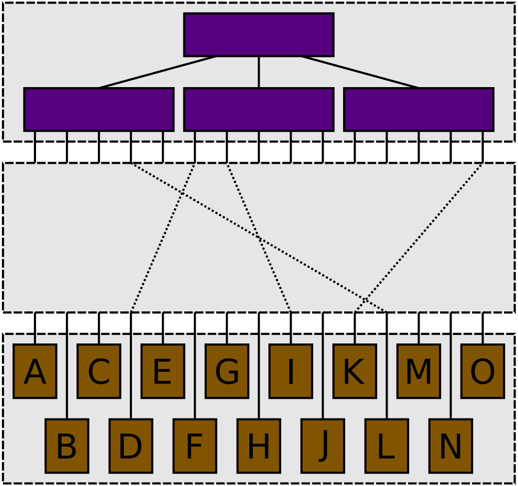
Optical Interconnects



Optical Interconnects

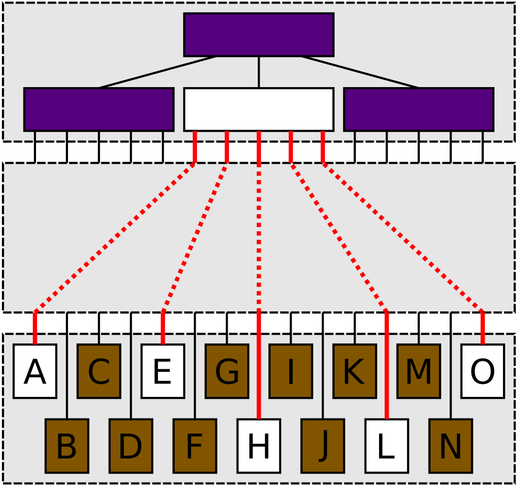


Groups stay consistent, but the communication pattern among members of the group can change rapidly



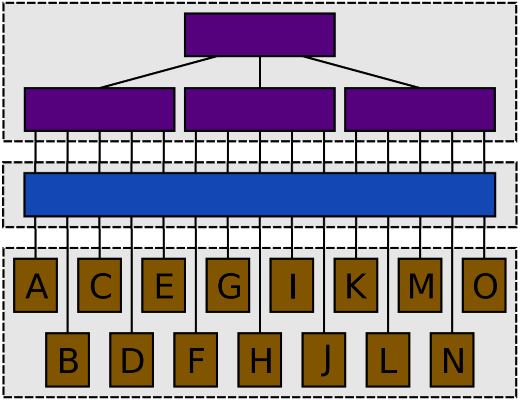
Use optical circuit switch to connect micro-servers to electrical switches

Group Membership - Example



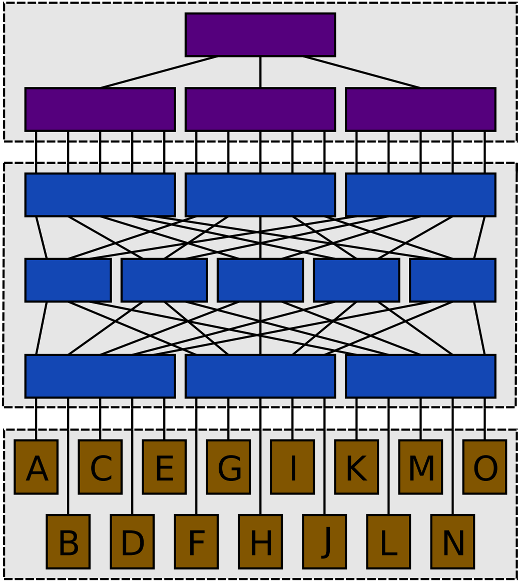
Allows the formation of arbitrary groups of micro-servers, when connectivity is required

Single Optical Circuit Switch



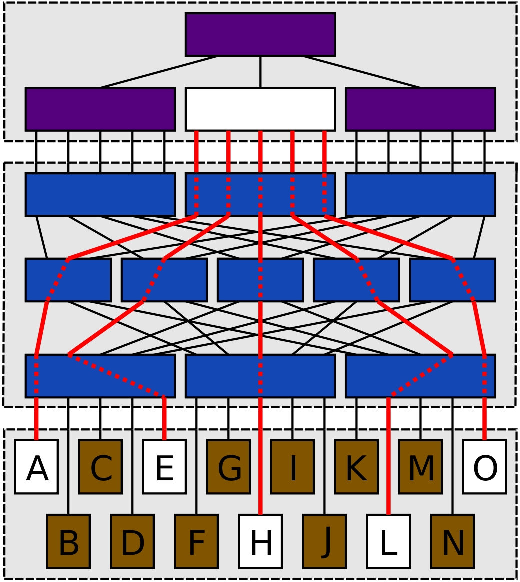
Optical circuit switches are not yet available beyond a few hundred ports

3 Stage Clos

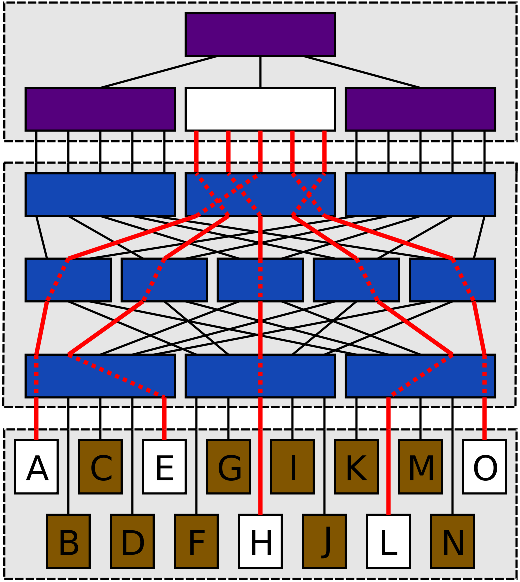


3 stage Clos provides the same functionality as a single switch

3 Stage Clos - Example

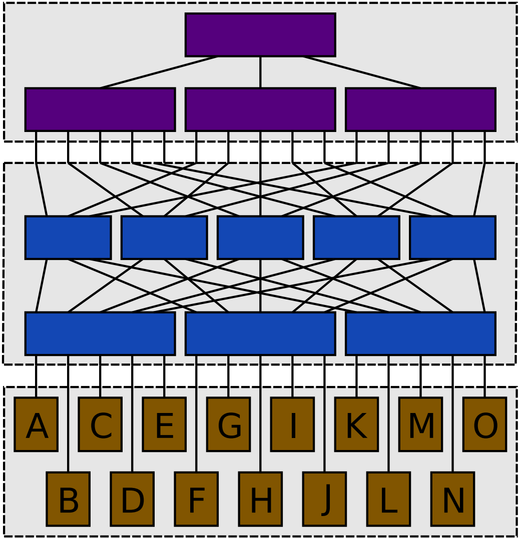


3 Stage Clos - Example



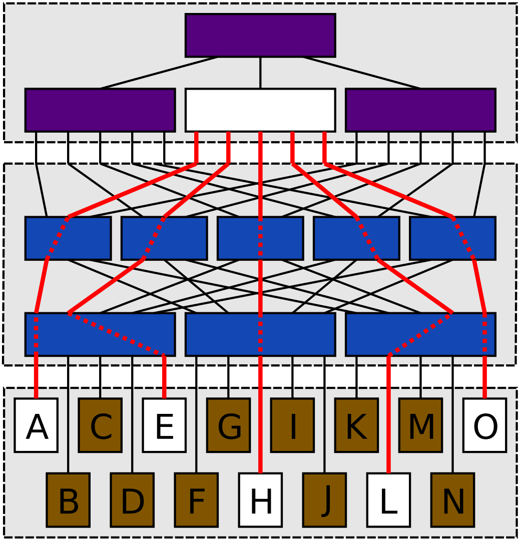
The exact port on the switch is not important

2 Stage Clos



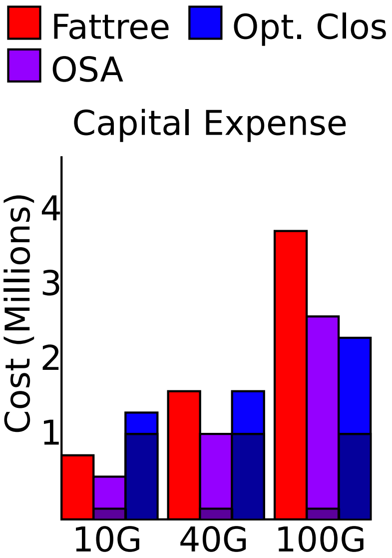
2 stage Clos provides sufficient flexibility to create any group

2 Stage Clos - Example



Any micro-server can reach any port on any switch, using 33% fewer optical ports than a 3 stage Clos

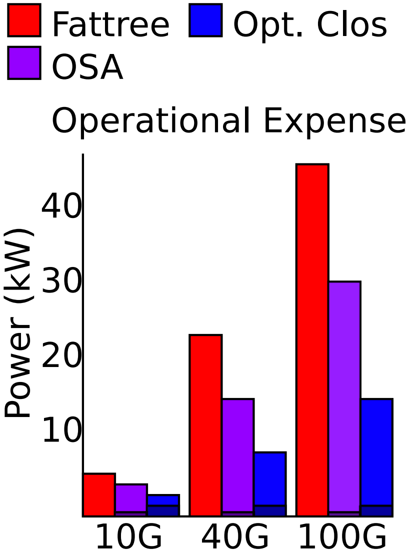
Cost Comparison



OSA requires less ports overall, and is the most cost effective for lower bandwidths

As the bandwidth moves into the 100s of Gbps, the cost of electrical switching dominates

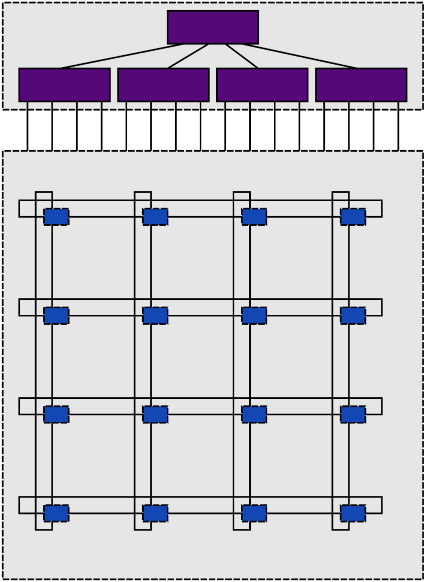
Power Comparison



Operating optical switches is substantially less power intensive than electrical switches

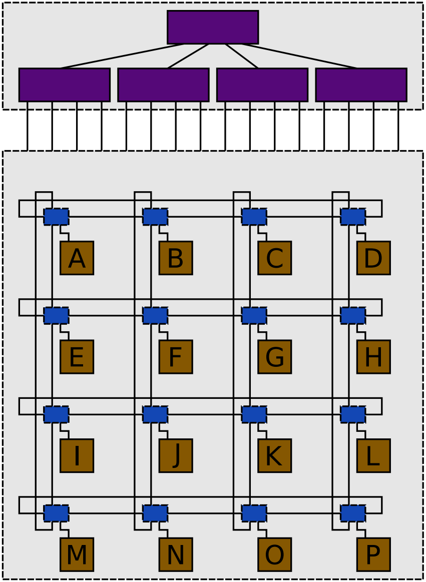
Green rack-scale computing must consider the impact of networking

Modular Circuit Switching



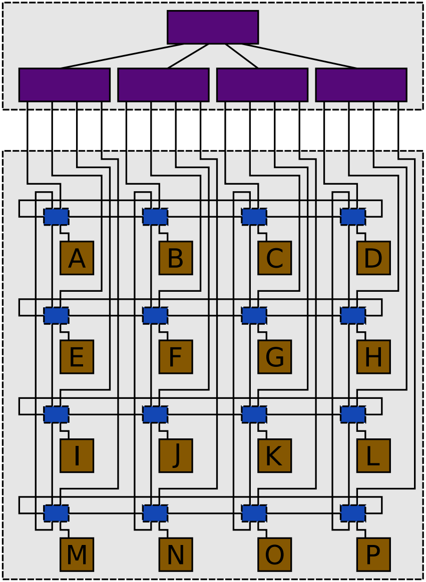
Perform circuit switching using a distributed set of circuit switches

Modular Circuit Switching



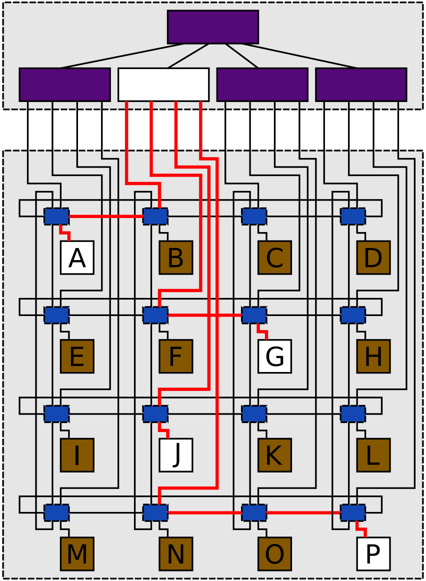
Each micro-server is connected to a switch

Modular Circuit Switching

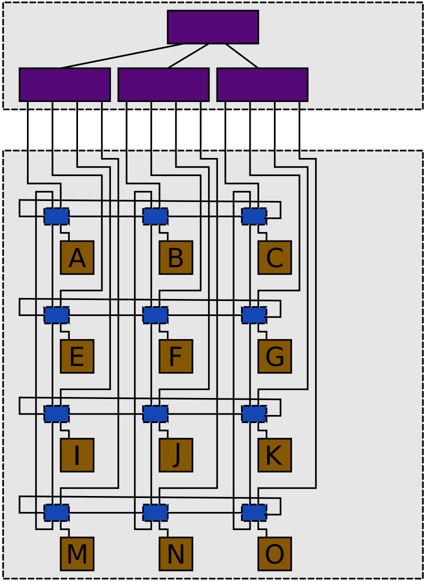


Each optical switch is connect to a port on an electrical switch

Modular Circuit Switching

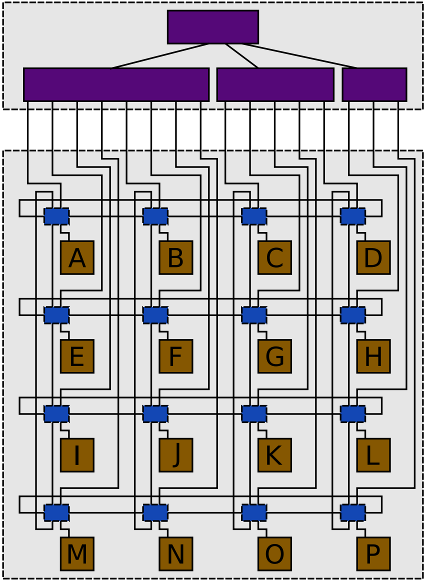


Modular Circuit Switching



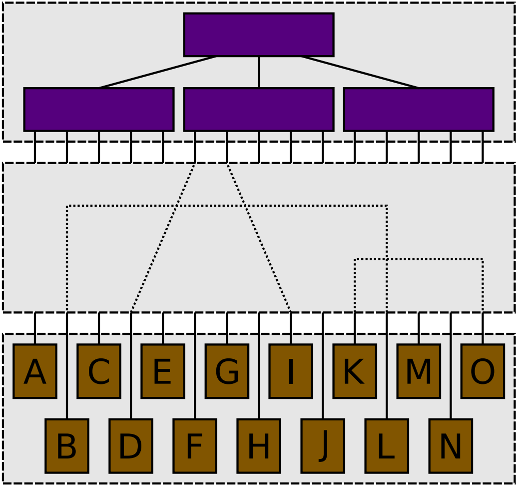
Only deploy the components that are needed

Modular Circuit Switching



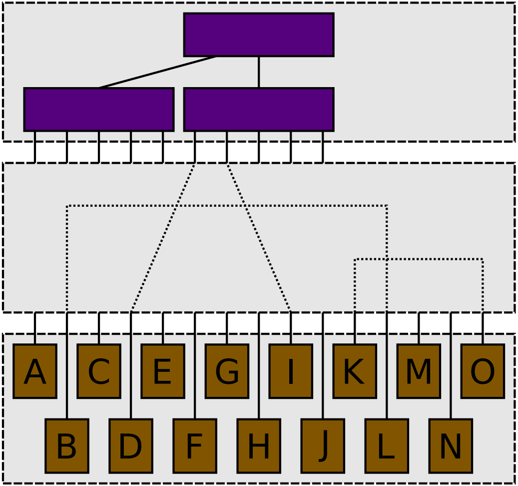
Supports various electrical switch sizes

Direct Connectivity



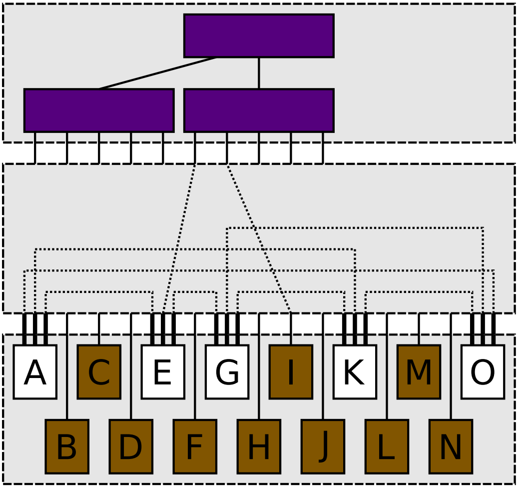
Can extend the concept to direct server to server connections

Direct Connectivity



Can eliminate some of the electrical switches

Direct Connectivity



Adding additional ports to micro-servers would allow dynamic construction of server centric networks

Summary

- What is the right network fabric for rack-scale computing?
- Data center networking solutions are not ideal at rack-scale
- We propose the use of reconfigurable optics to form groups
- The idea extends to dynamically constructing other topologies