FUJITSU

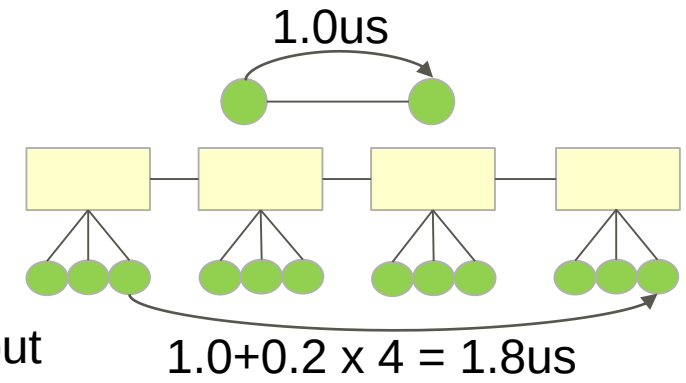
shaping tomorrow with you

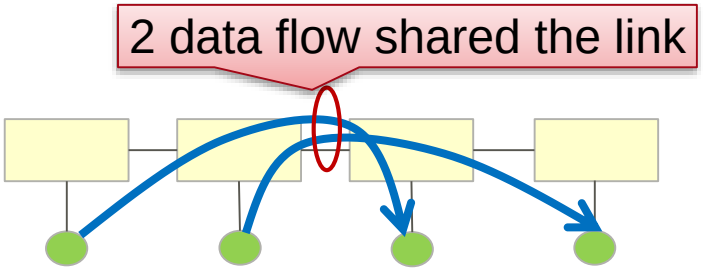
# Conservative Interconnect of Large-scale HPC systems

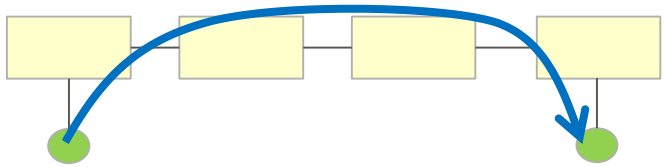
Kohta Nakashima
Fujitsu Laboratories Ltd.
2019.11.26

# Introduction

- Interconnect technologies are important for HPC systems
- For performance
  - Low latency and high bandwidth can improve system performance
  - Latency (InfiniBand):
    - Direct connection: 1us
    - Switch latency: 0.2us/switch
  - Throughput:
    - Average shortest path length affects throughput



- For cost
  - Small # of switches and links can reduce system cost

Smaller diameter and average shortest path length improve the HPC system performance and cost

# Outline

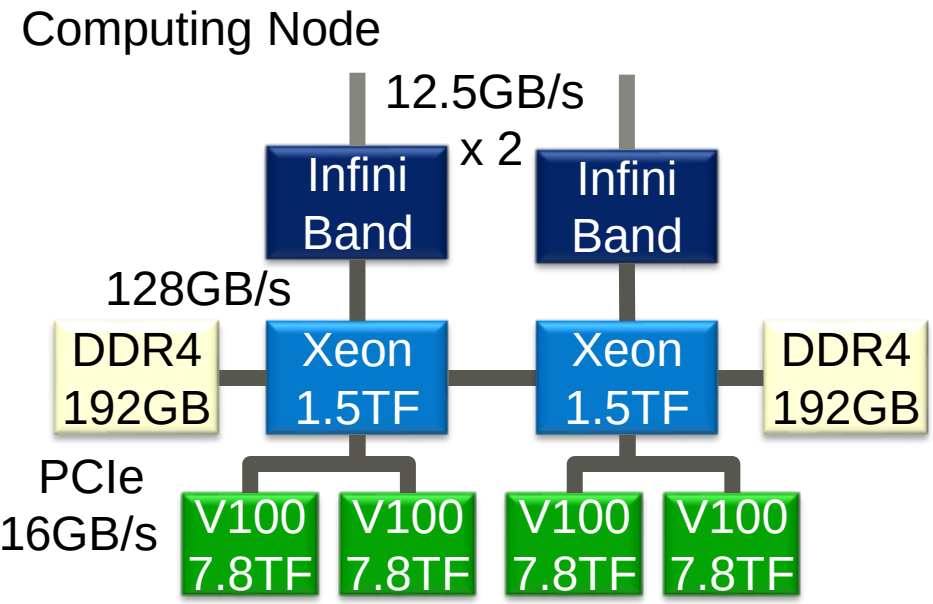
FUJITSU

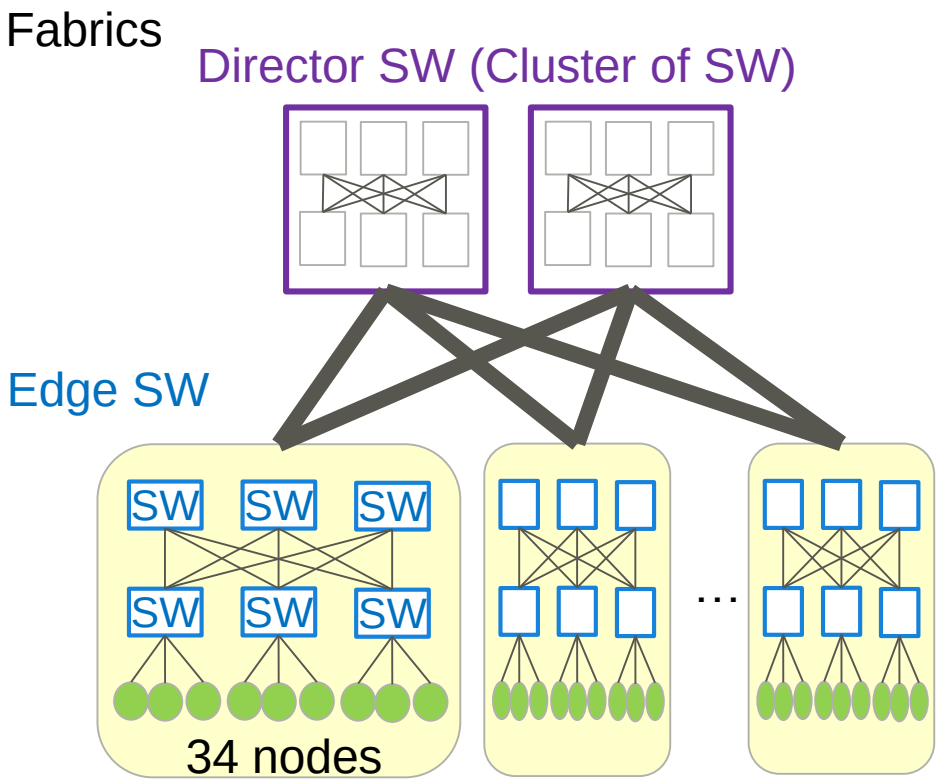
- Introduction
- Example of HPC systems
- Network topologies for HPC systems
- The reason why the HPC networks are so conservative
- How to explorer to innovate HPC networks

# Example of supercomputer in Japan (1)

FUJITSU

- ABCI: AI Bridging Cloud Infrastructure (2018)
  - # of nodes: 1,088
  - Peak Flops: 37PFlops
  - HPL perf.: 19.88PF
    (5th in Top500, 2018/6)
  - NVIDIA Tesla V100
  - InfiniBand x 2

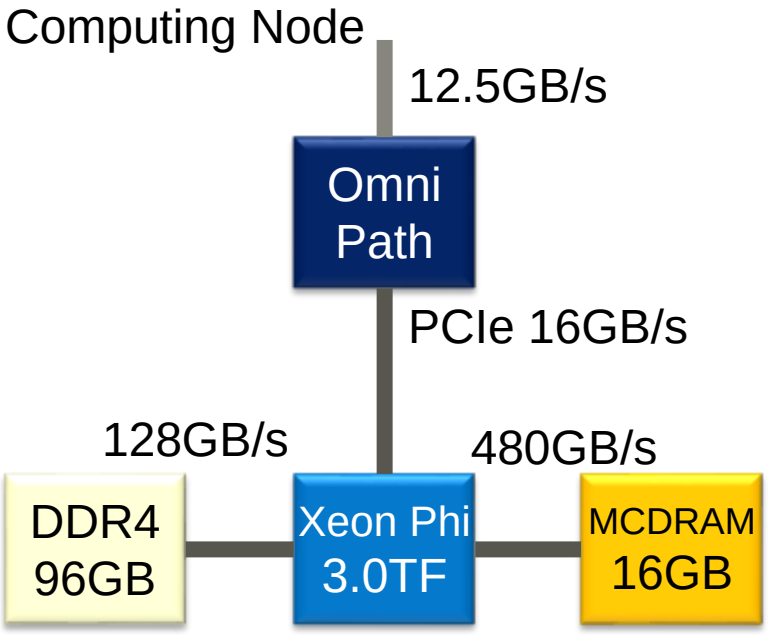


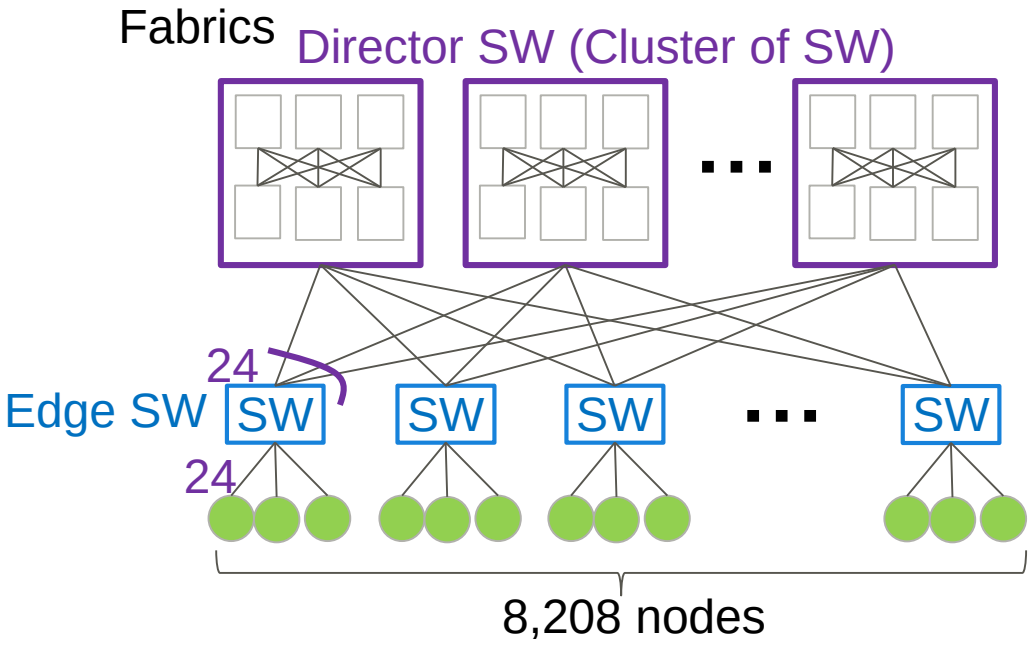



# Example of supercomputer in Japan (2)

FUJITSU

- Oakforest-PACS (2016)
  - # of nodes: 8,208
  - Peak Flops: 25PFlops
  - HPL perf.: 13.55PF
    (6th in Top500, 2016/11)
  - Xeon Phi 68 cores/1.4GHz
  - OmniPath 100Gbps(12.5GB/s)

# 20 systems in Top500 list

FUJITSU

- Topology
  - Fat Tree: 16 systems
  - Dragonfly: 3 systems
  - Torus: 1 system
- Interconnect
  - InfiniBand: 7 systems
  - OmniPath: 5 systems
  - Aries(Cray): 3 systems
  - TX2: 1, BXI: 1, Custom: 2

| Rank | Name | Fabric |
|---|---|---|
| 1 | Summit | Fat Tree/IB |
| 2 | Sierra | Fat Tree/IB |
| 3 | TaihuLight | Fat Tree/Custom |
| 4 | Tianhe-2A | Fat Tree/TX2 |
| 5 | Frontera | Fat Tree/IB |
| 6 | Piz Daint | DF/Aries |
| 7 | Trinity | DF/Aries |
| 8 | ABCI | Fat Tree/IB |
| 9 | SuperMUC-NG | Fat Tree/OPA |
| 10 | Lassen | Fat Tree?/IB |

| Rank | Name | Fabric |
|---|---|---|
| 11 | PANGEA III | Fat Tree?/IB |
| 12 | Sequoia | Torus/Custom |
| 13 | Cori | DF/Aries |
| 14 | Nurion | Fat Tree/OPA |
| 15 | Oakforest-PACS | Fat Tree/OPA |
| 16 | HPC4 | Fat Tree?/IB |
| 17 | Tera-1000-2 | Fat Tree/BXI |
| 18 | Stampede2 | Fat Tree/OPA |
| 19 | Marconi | Fat Tree/OPA |
| 20 | DGX SuperPOD | Fat Tree?/IB |

DF: Dragonfly, IB: InfiniBand, TX2: TH Express 2, OPA: OmniPath, BXI: Bull Exascale Interconnect

# Other topologies

FUJITSU

- 16 Fat Tree, 3 Dragonfly, 1 Torus in Top20
- Others
  - Dragonfly+
    - Gadi, National Computational Infrastructure (NCI Australia), #47
    - Niagara, University of Toronto, #76
  - Hypercube
    - Eagle, National Renewable Energy Laboratory, 8D Hypercube #43

- Top500: Fat Tree, Dragonfly, Torus, Dragonfly+, Hypercube
  - Fat Tree: Almost all systems
  - Dragonfly: Cray users
  - Torus: Fujitsu FX series, IBM BlueGene, and Sugon users
  - Dragonfly+: Challenging users with Mellanox InfiniBand
  - Hypercube: A part of SGI users

# Fabrics

FUJITSU

- Open
  - Third party can purchase it and integrate the system using the fabrics
  - InfiniBand (Mellanox) and OmniPath (Intel)

| Fabric | Vendor | Topology |
|---|---|---|
| InfiniBand | Mellanox | Fat Tree, Dragonfly+, Torus, Hypercube, etc |
| OmniPath | Intel | Fat Tree |

- Custom
  - Vendor only provides the system integrated by the fabrics

| Fabric | Vendor | Topology |
|---|---|---|
| Aries/Slingshot | Cray | Dragonfly |
| BXI (*) | Atos | Fat Tree |
| Tofu | Fujitsu | Torus |

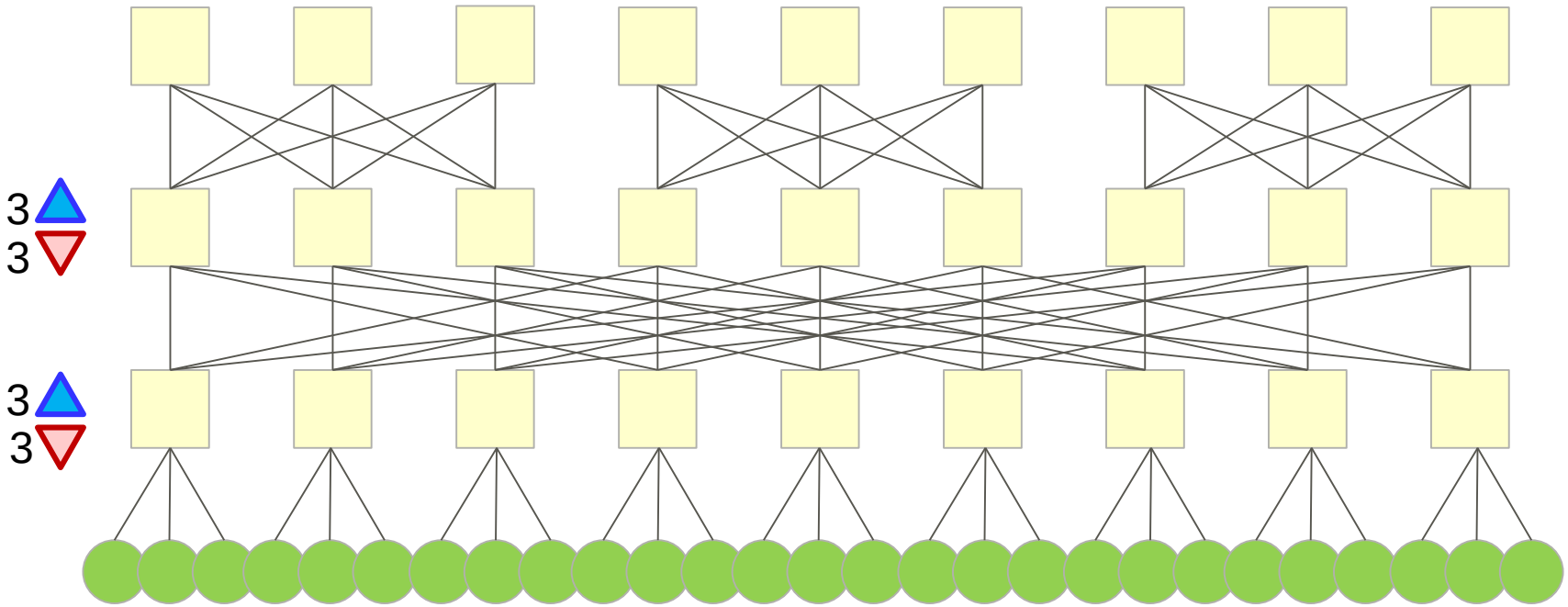
(* BXI: Bull eXascale Interconnect, Bull was purchased by Atos)

# Outline

- Introduction
- Example of HPC systems
- Network topologies for HPC systems
- The reason why the HPC networks are so conservative
- How to explorer to innovate HPC networks

# Fat Tree

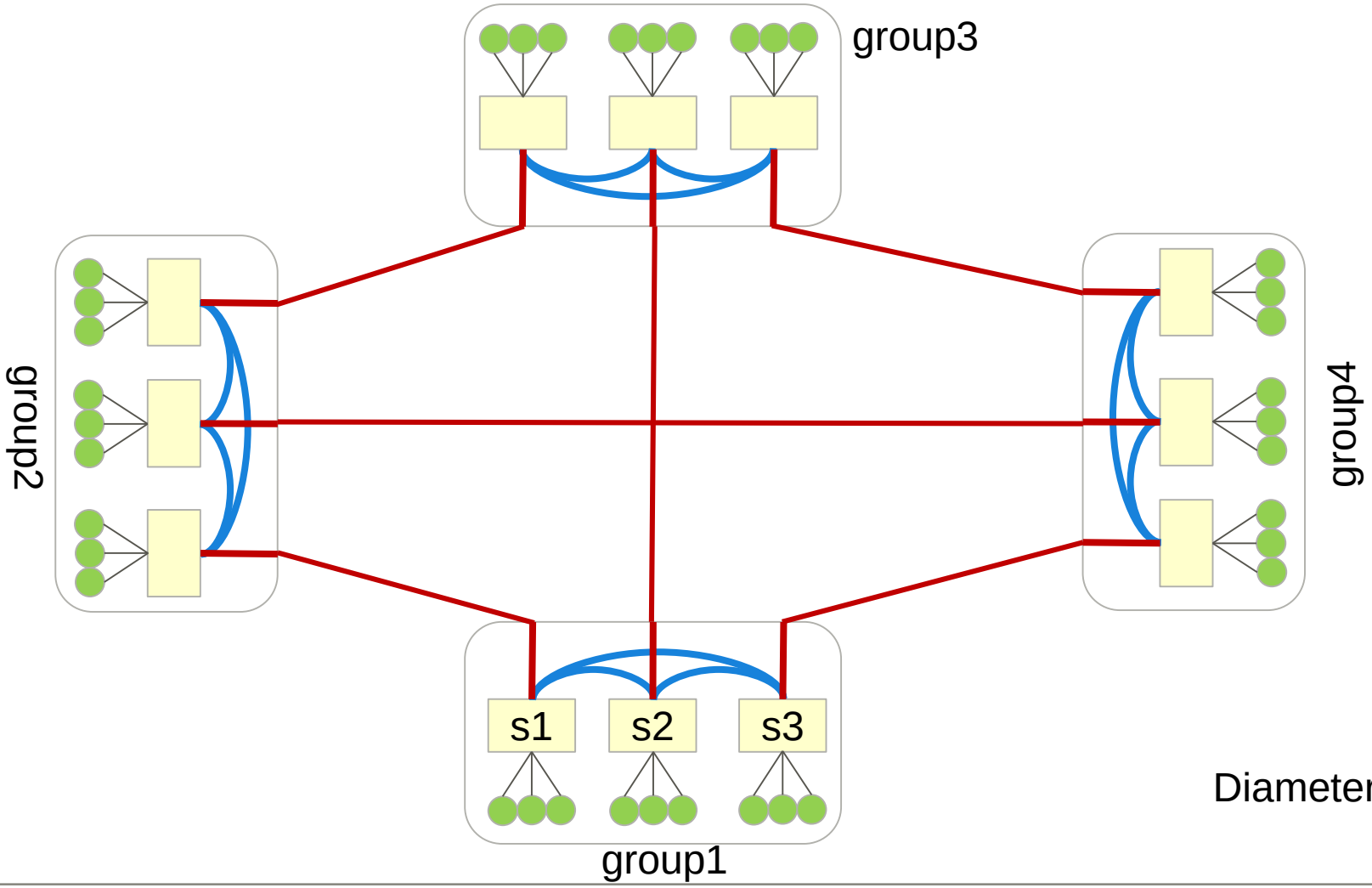
- Non blocking, if # of ports for uplink and downlink is same
- Configurable for performance, cost




Diameter: 6

# Dragonfly

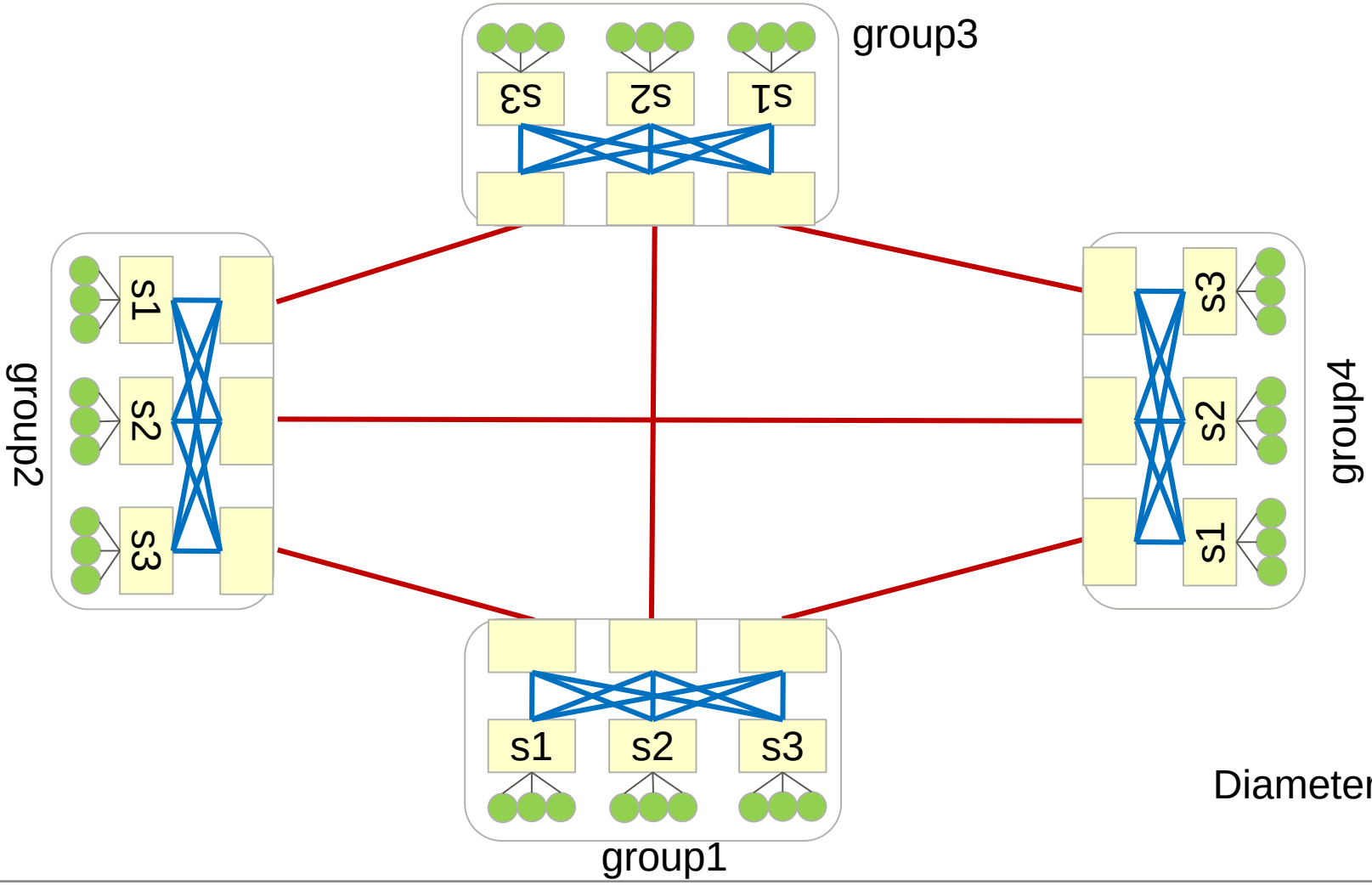
- Local: Local switches are connected directly in group
- Global: Groups are connected directly




Diameter: 5

# Dragonfly+

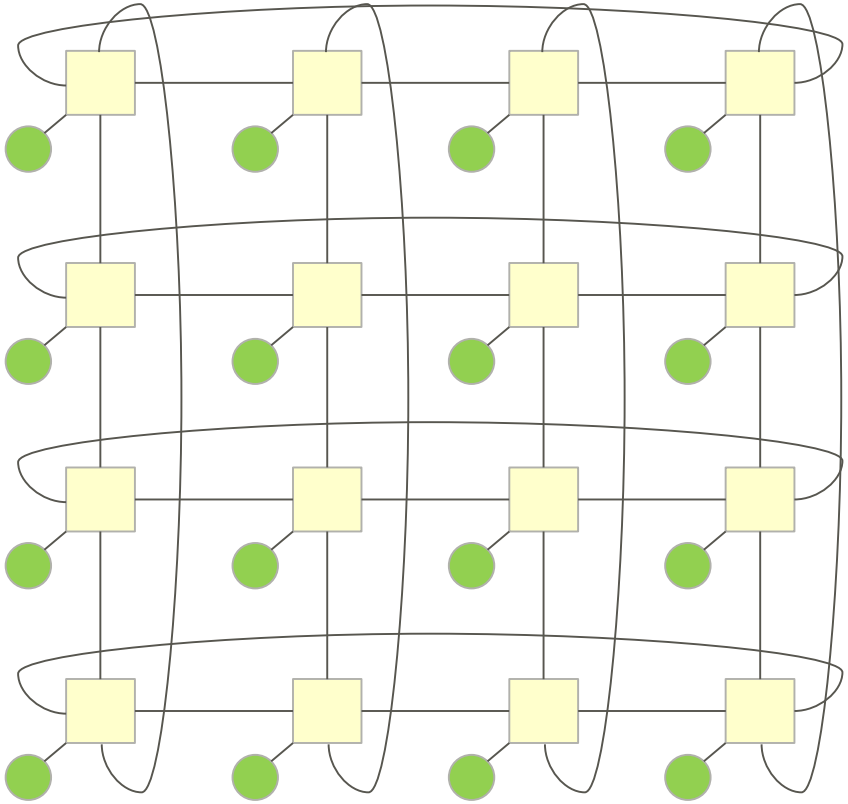
- Local: Local nodes are connected by Clos (like Fat tree)
- Global: Groups are connected directly




Diameter: 5

# Torus

- Scalable, easy to connect physically for 10,000+ nodes



2D Torus

# My recommendation for customers/colleagues

FUJITSU

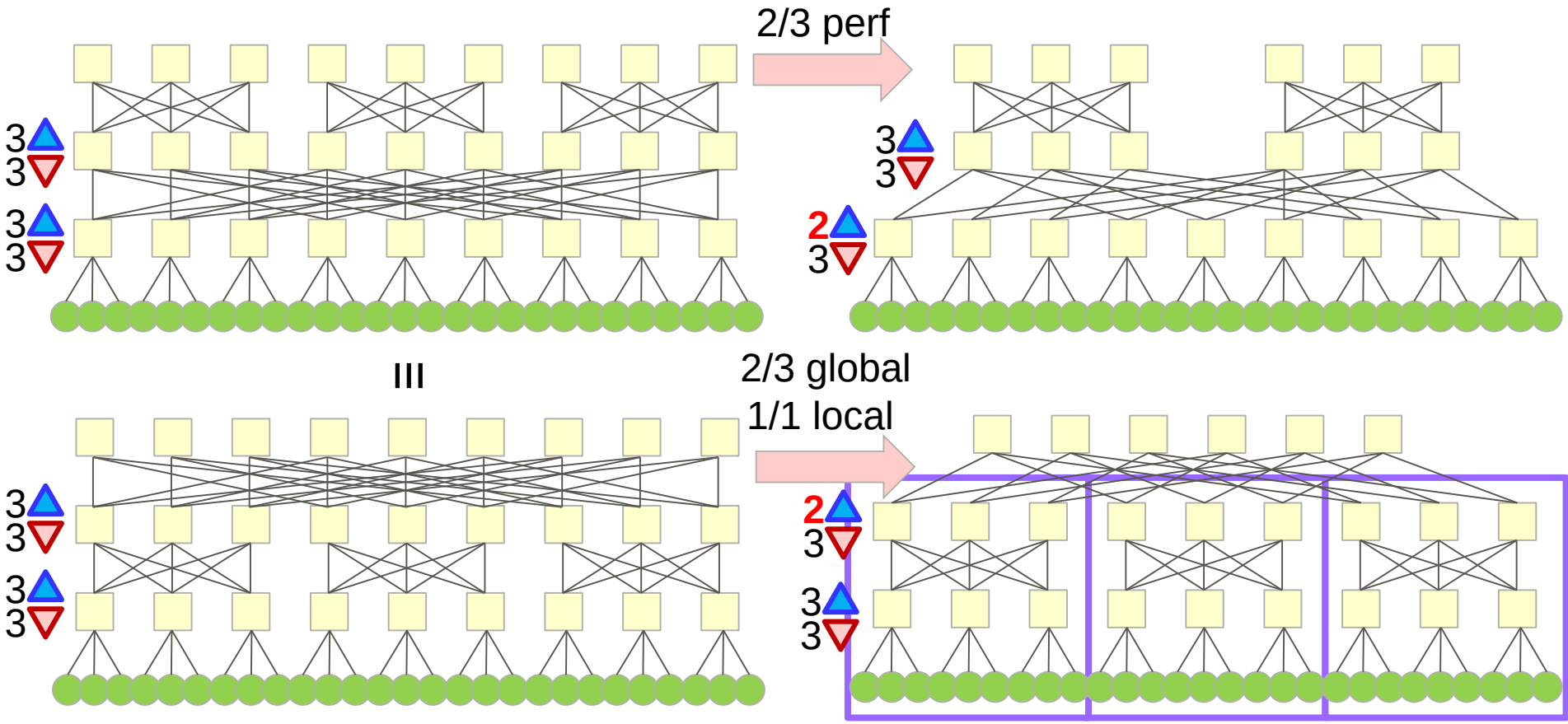
- > 16,000 nodes or Fujitsu FX series customer
  - Torus
  - Easy for physical installation
- 16,000 ~ 800
  - For major customer
    - 3-level Fat Tree
    - Mature
  - For challenging customer
    - DF/DF+
    - DF+: Mellanox support
- < 800 nodes
  - 2-level Fat Tree
  - Perfect

Who is rival?

# Detail of Fat Tree features (1)

FUJITSU

- Easy to configure balance between performance and cost
- Easy to explain the configuration of fabric

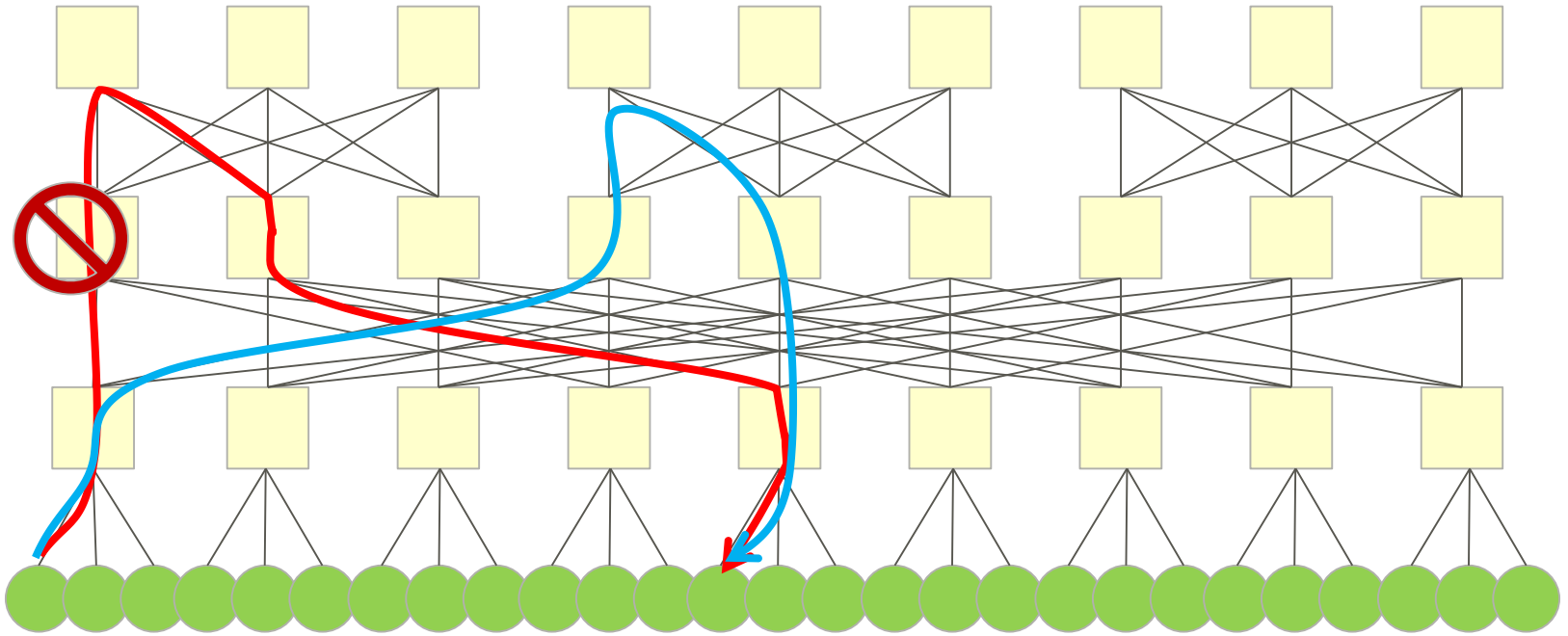



Sales engineer can configure the fabric for customer easily

# Detail of Fat Tree features (2)

FUJITSU

- Easy to explain the behavior if some links or switches failed



# Outline

- Introduction
- Example of HPC systems
- Network topologies for HPC systems
- The reason why the HPC networks are so conservative
- How to explorer to innovate HPC networks

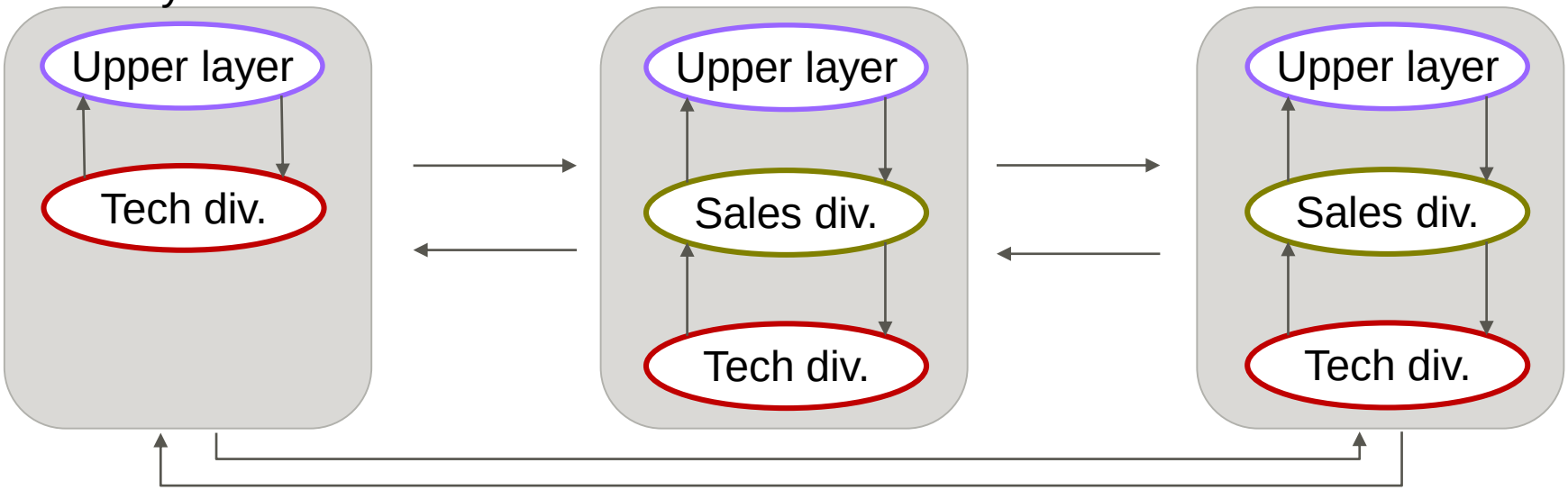
# Road to decide configuration of HPC systems

FUJITSU

System owner
- University/National lab
- Industry

System vendor
- IBM/Cray/Fujitsu etc

Component vender
- Intel/AMD/NVIDIA etc
- Mellanox/Intel etc




- Define specification for the procurement
- Explain suitable purpose of the system

- Propose the system design
- Explain whether the procurement is good for business

- Consult and promote components

Important to convince various stakeholders

# Why the HPC networks are so conservative

FUJITSU

- Because Fat Tree is very suitable for HPC systems
  - Easy to explain for the design
  - Easy to configure cost and performance
  - A little expensive but mature

- Because stakeholders are conservative
  - Hard to challenge the novel technologies with high risk because total budget of system is very high (1~1,000 M$)
  - Stakeholders
    - System owners
    - System vendors
    - Component venders

# Outline

- Introduction
- Example of HPC systems
- Network topologies for HPC systems
- The reason why the HPC networks are so conservative
- How to explorer to innovate HPC networks

# Conditions for acceptable topologies (1)

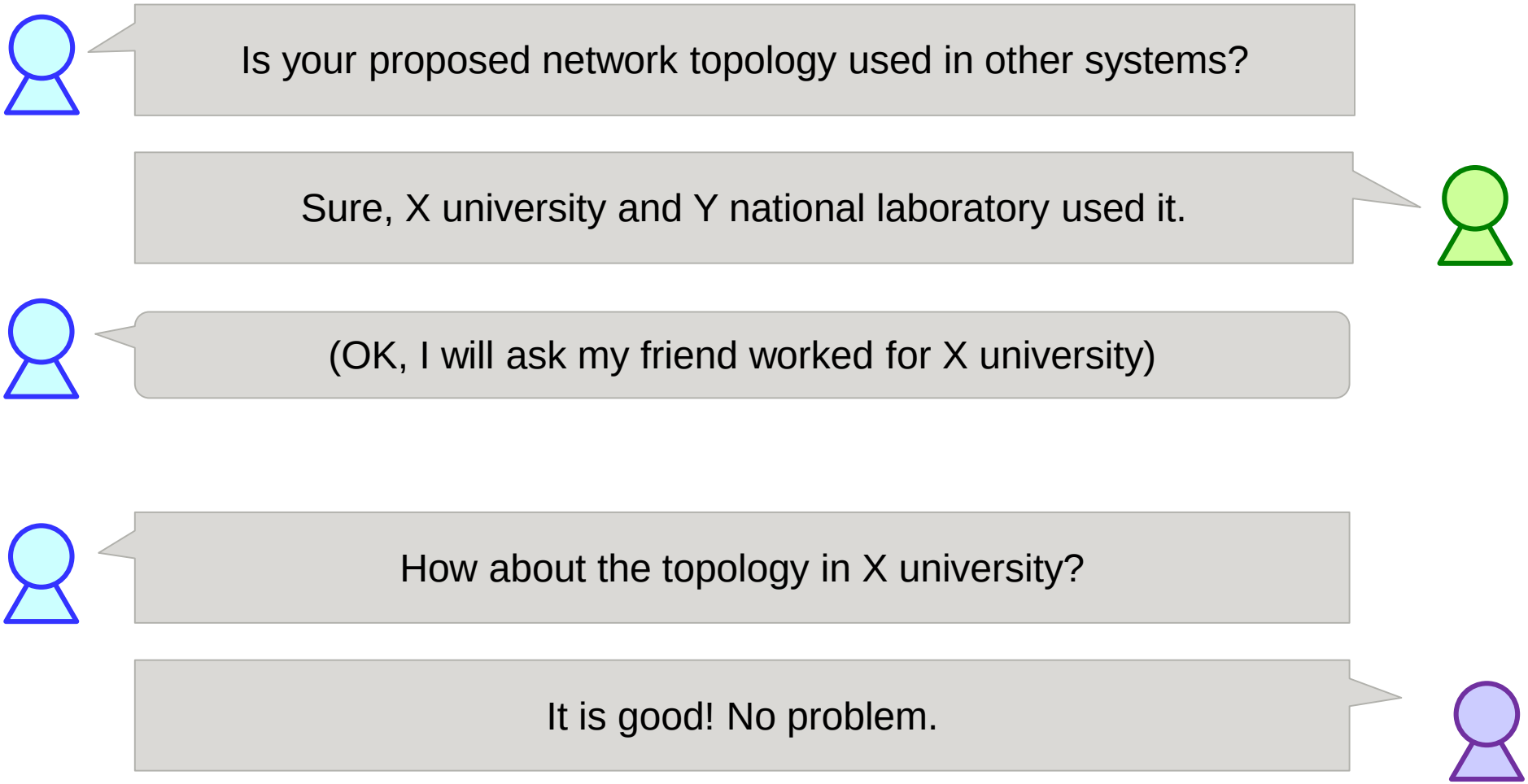
FUJITSU

- Easy to understand
- Easy to configure the performance and cost
  - For not only engineers but also sales division
  - Without simulation, only use paper and pencil
- Feel low risks
  - Simple logic for routing, deadlock avoidance, and fault tolerant
  - Easy to find another routing if some links failed
- Feel relieved somehow…

# Conditions for acceptable topologies (2)

FUJITSU

- Proven/Mature
  - Easy to convince if other systems use the topology



# How to make stakeholders adopt novel topology? (1)

FUJITSU

- Strategy 1：Collaboration with interconnect vendor
- Pros
  - If the topology has strong advantage compared to other topology, it can motivate interconnect vender to support.
  - If interconnect vendor support the topology, system owners and system venders feel low risk and relieved
- Cons
  - If the topology reduce # of switches and/or links, it may be unacceptable for interconnect vender
  - Almost all good topology can reduce # of switch; hard to accept to support it for interconnect vendor
- Example of success：Dragonfly+ by Mellanox

# How to make stakeholders adopt novel topology? (2)

FUJITSU

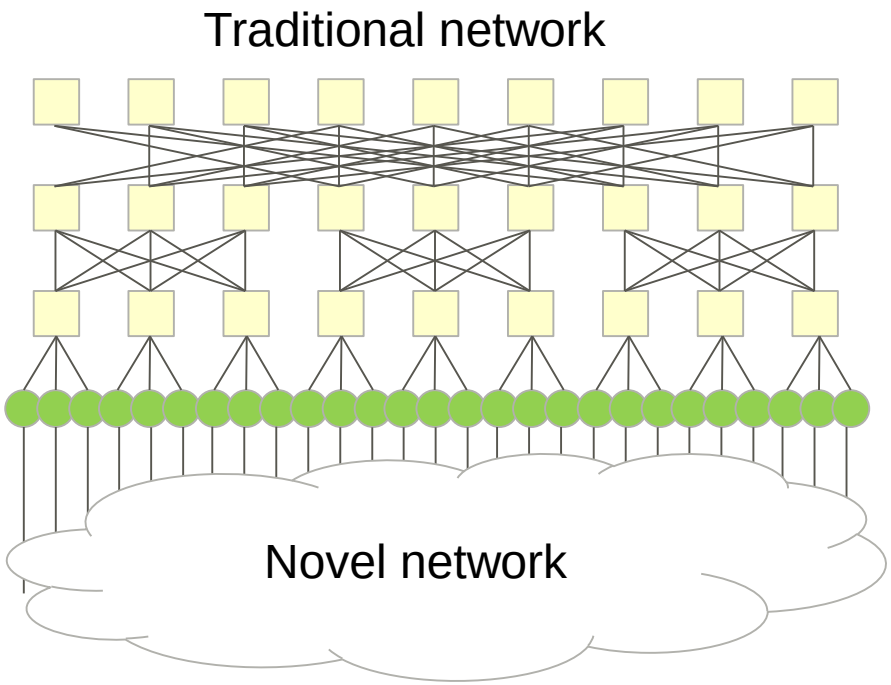
- Strategy 2 ：Verification of usefulness of the topology through research collaboration with understanding customers
- Pros
  - Customers who are also research institutions jointly verify usefulness, so they are easy to accept
  - If the customers are authority of interconnect, other customers may accept the topology
- Cons
  - Requires the system for verification for research collaboration
    - The system requires 100s computing nodes and switches…
    - Total cost: at least several million dollars…
  - The system cost of verification:
    - For customer: Too high. Because the verification may fail
    - For system vendor: Too high as an investment, other customer may not accept the topology even if the verification succeeded

# How to make stakeholders adopt novel topology? (3)

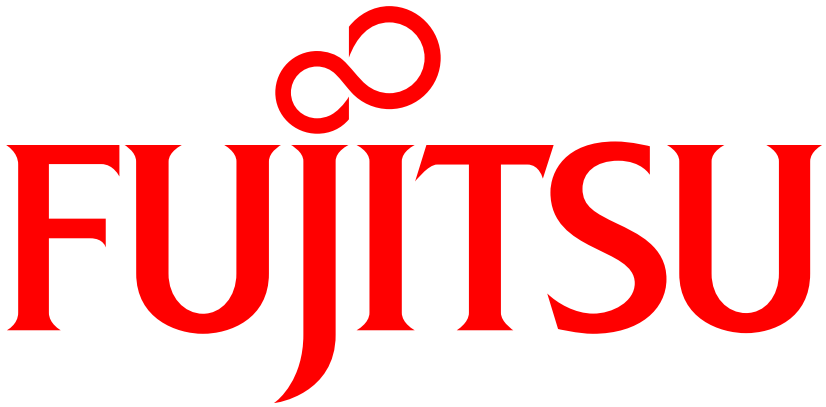
- Strategy 3：National project for supercomputing
- Pros
  - High probability of success because of huge investment
- Cons
  - Unacceptable too high risk technologies because of huge investment
  - If the technologies for the project focus only huge systems, the merit of the technologies for mid-range systems may decrease…
- Example of success：
  - Dragonfly：XC30(Cascade), HPCS Project in DARPA, USA
  - 6D-Mesh/Torus: Tofu – K computer, National project for HPC in Japan

# Summary and proposal

- Summary: HPC interconnect is very conservative
  - Fat tree is very suitable for practical HPC systems
  - Stakeholders do not like take a risk




- Proposal
  - Network combination
    - Traditional network guarantees the minimum performance and fault tolerance
    - Novel network performs excellent performance
  - Additional evaluation value for the novel network
    - Diameter and average shortest path length with failure switches and links
    - Worst case results when n switches/links failed




shaping tomorrow with you