FUJITSU

shaping tomorrow with you

富士通のPCクラスタへの取り組み

～PCクラスタ向け次世代トポロジー実現に向けて～

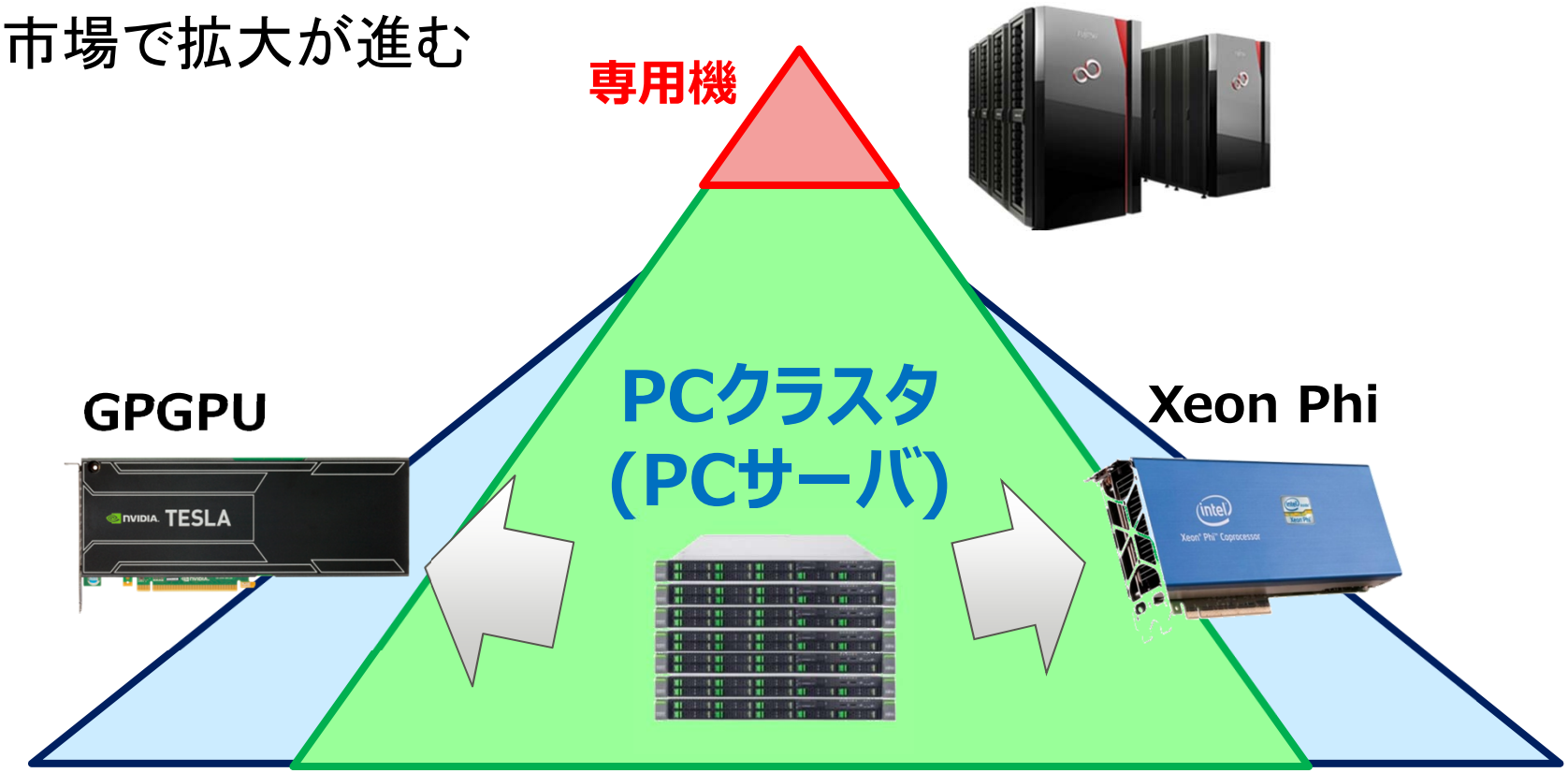
2014.12.12

(株)富士通研究所
中島 耕太
nakashima.kouta@jp.fujitsu.com

アクセラレータによる高速化

FUJITSU

- アクセラレータを搭載したPCクラスタがHPCの主流
  - 専用機と比較してコストパフォーマンスに優れる
  - 市場で拡大が進む



アクセラレータの活用により演算性能が大幅に増強

ネットワーク性能

FUJITSU

- アクセラレータの広がり
  - GPGPUやXeon Phiといったアクセラレータの利用によりノードあたりの演算性能が大幅に増加
  - 演算性能の向上に合わせてネットワーク性能も高めたい
- 演算とネットワークの性能バランス

ノードあたり性能	2008年 CPU中心			2013年 アクセラレータ中心		
	T2K筑波	T2K東大	T2K京大	HA-PACS	TSUBAME	名大
演算性能(GFlops)	147	147	147	2,993	4,071	1,521
ネットワーク(GB/s)	8GB/s	5GB/s	8GB/s	8GB/s	8GB/s	7GB/s
比率	0.054	0.034	0.054	0.003	0.002	0.005

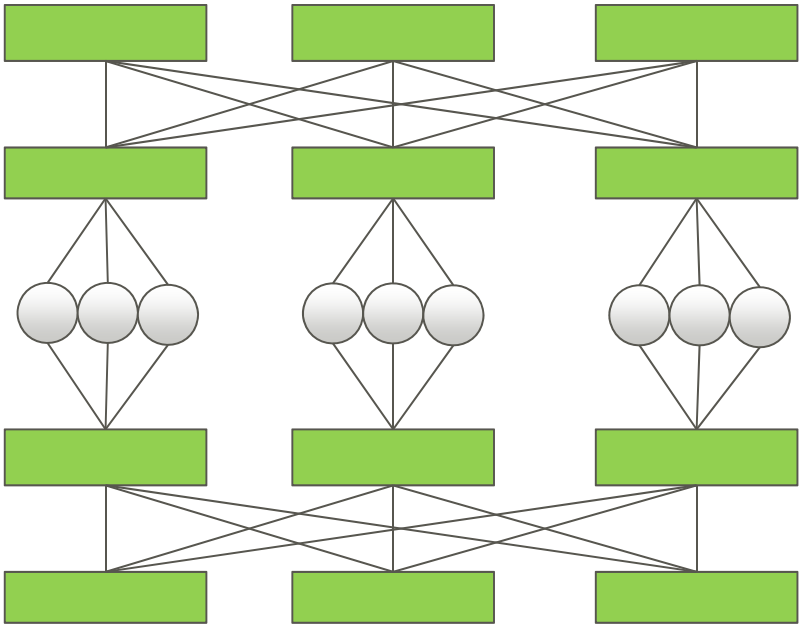
大きく低下

相対的にネットワーク性能が低下

ネットワーク性能を向上させるには

FUJITSU

- 1本あたりの帯域を増やす
- ネットワークの多重度を増やす



スイッチ数増加に伴う様々なコスト増加

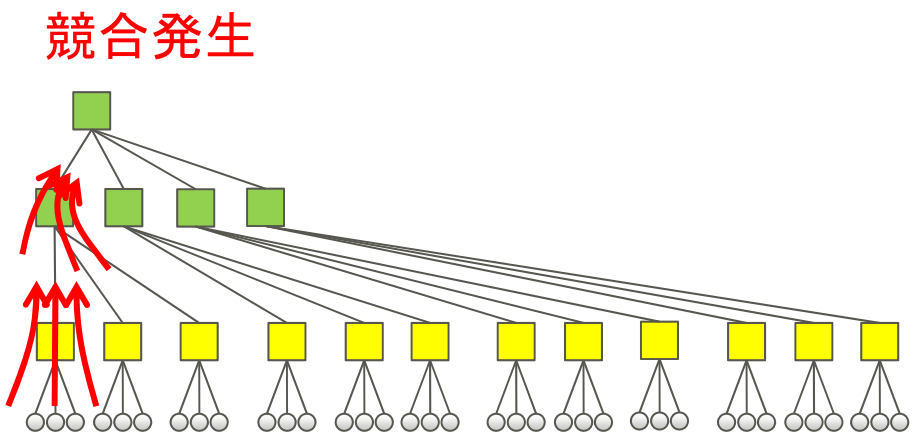
部材　電力　設置面積

より少ないスイッチ数で高い性能を実現できるネットワークが必要

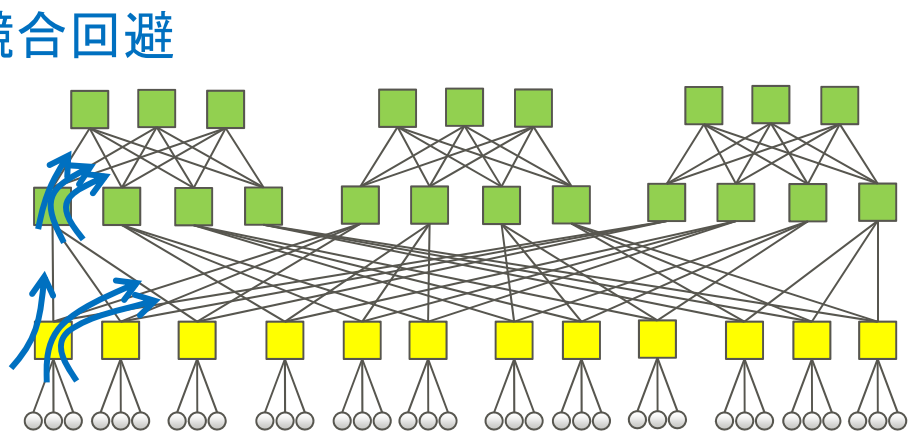
Fat Treeの利点と欠点

FUJITSU

- Fat Tree: PCクラスタでもっとも広く用いられているトポロジー



(Fat ではない) Tree



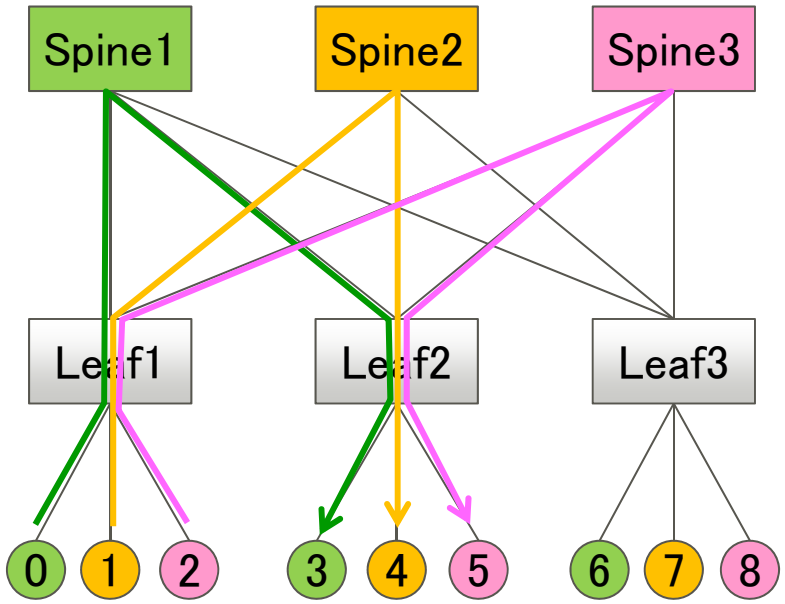
Fat Tree

- 利点: 高性能
  - 通信負荷が高いAll-to-all通信でも高い性能を実現
- 欠点: スイッチ数が多い
  - 電力、設置面積、部材のコストが高い
  - コスト削減のために帯域を絞ると性能が大きく劣化

IBにおけるFat Treeの経路制御とAll-to-all通信

FUJITSU

- 経路制御: 転送先ノードによって経由するSpineを切り替える
- Alltoall: すべてのノードのデータを全てのノードに渡す



各フェーズでの送信先 \ 送信元	Leaf1			Leaf2			Leaf3		
	0	1	2	3	4	5	6	7	8
0	0	1	2	3	4	5	6	7	8
1	1	2	3	4	5	6	7	8	0
2	2	3	4	5	6	7	8	0	1
3	3	4	5	6	7	8	0	1	2
4	4	5	6	7	8	0	1	2	3
5	5	6	7	8	0	1	2	3	4
6	6	7	8	0	1	2	3	4	5
7	7	8	0	1	2	3	4	5	6
8	8	0	1	2	3	4	5	6	7

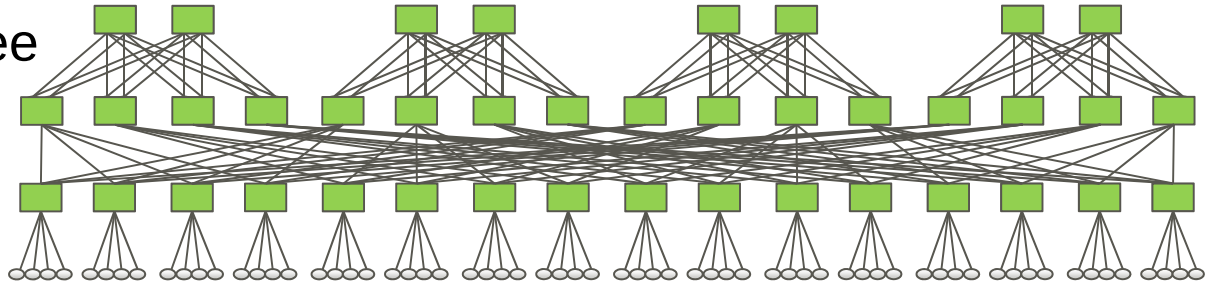
各フェーズにおいて経路競合が回避できており高性能

Fat Treeが必要とするスイッチ数

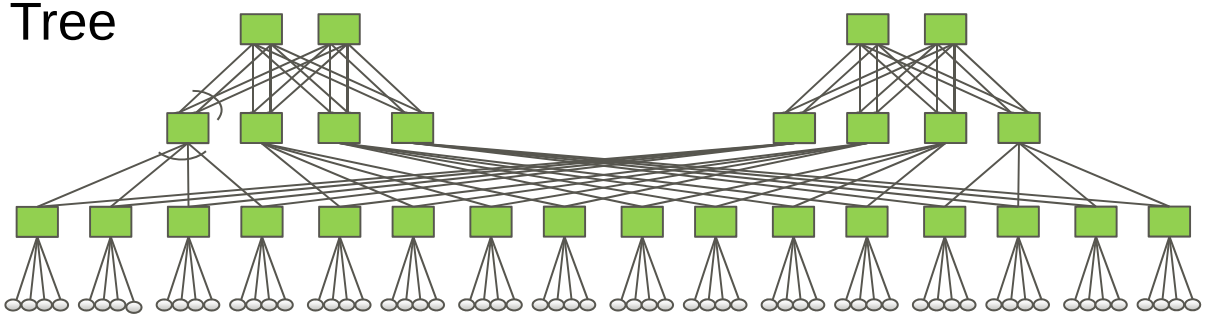
FUJITSU

■6,000ノードクラスのFat Treeを比較

FBB Fat Tree



1/2帯域Fat Tree



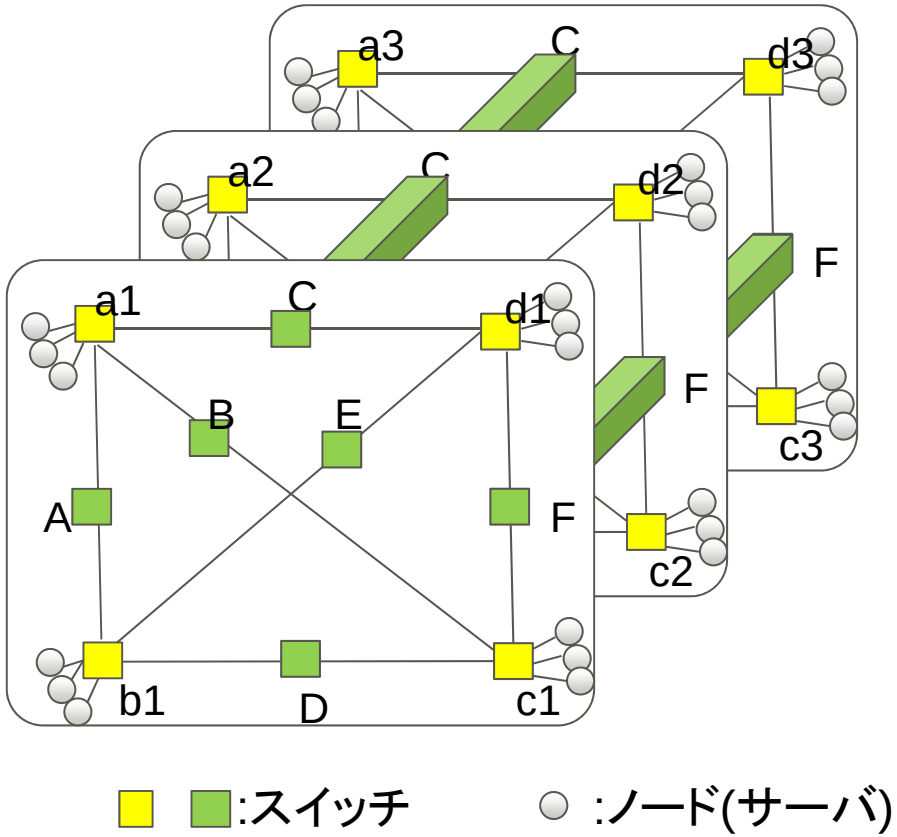
	ノード数	スイッチ数
FBB Fat Tree	5,832	810
1/2 Fat Tree	5,832	567

810台のスイッチが必要

高い性能とスイッチ数削減を両立するために

FUJITSU

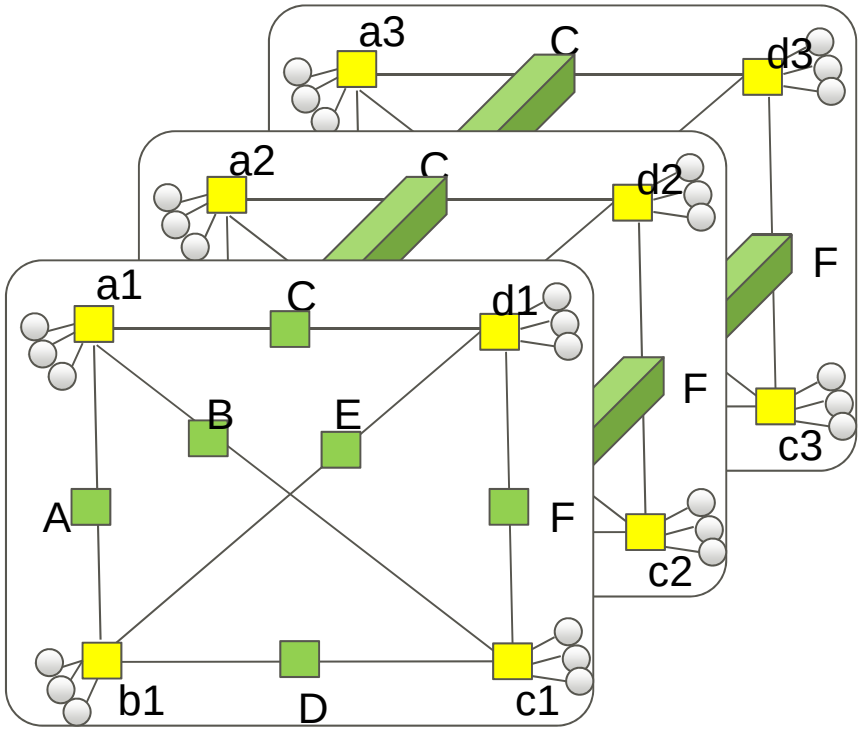
- 多層Fullmeshトポロジーを提案
  - 3段Fat Treeと比較しスイッチ数を4割削減
  - データ交換順序の工夫で**All-to-all通信**性能をFat Treeと同一に



必要とするスイッチ数

FUJITSU

■6,000ノードクラスを構成を考える



(上図は4角形3層で36ノード、下表構成は、19角形18面でノード)

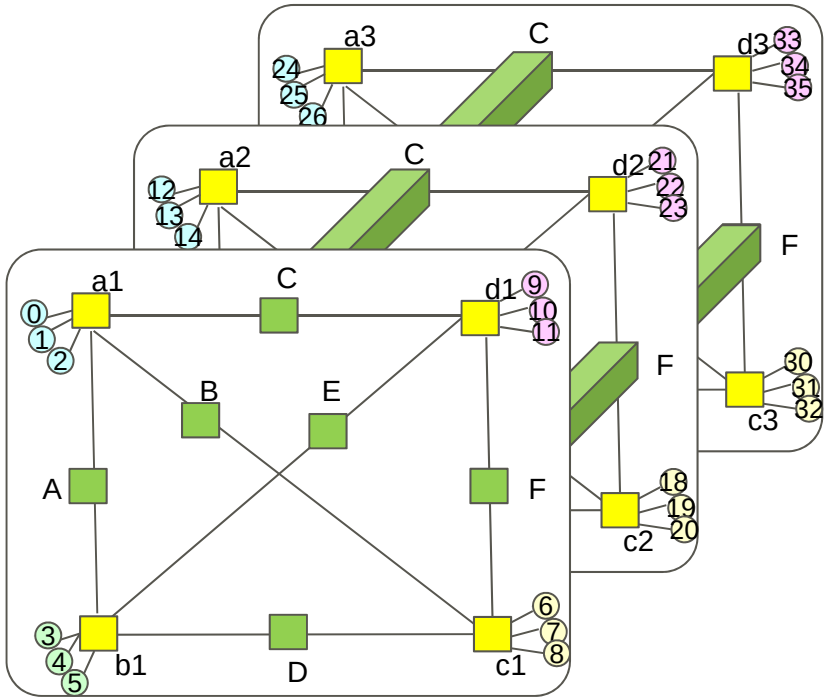
	ノード数	スイッチ数
FBB Fat Tree	5,832	810
1/2 Fat Tree	5,832	567
多層Fullmesh	6,156	513

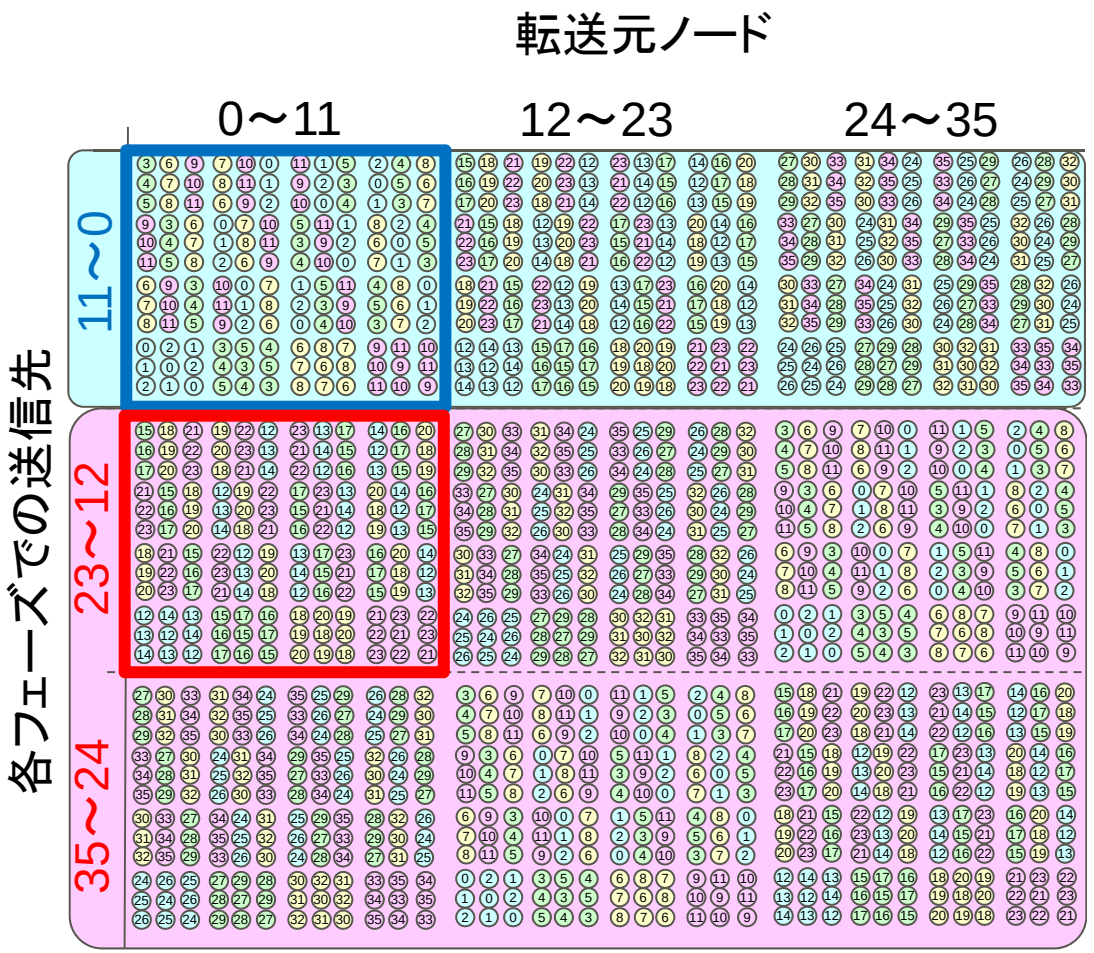
4割削減

多層Fullmeshにおける All-to-all通信

FUJITSU

■36ノードが36フェーズに分けて全ノードに送信

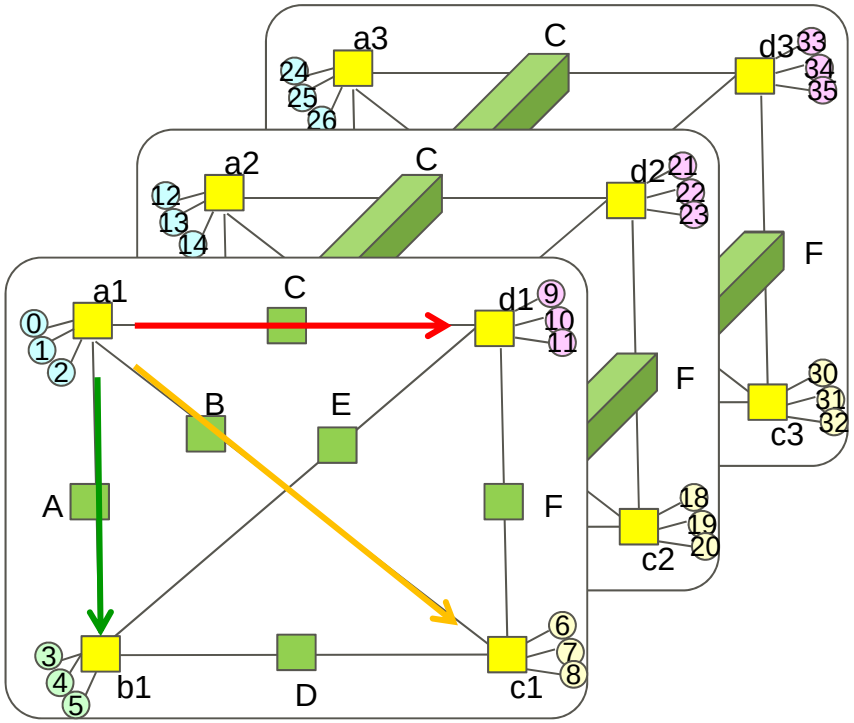




層内に閉じる通信と層を跨ぐ通信についてそれぞれ詳細を説明

層内に閉じる通信

■別々の頂点に向かうようにスケジュール

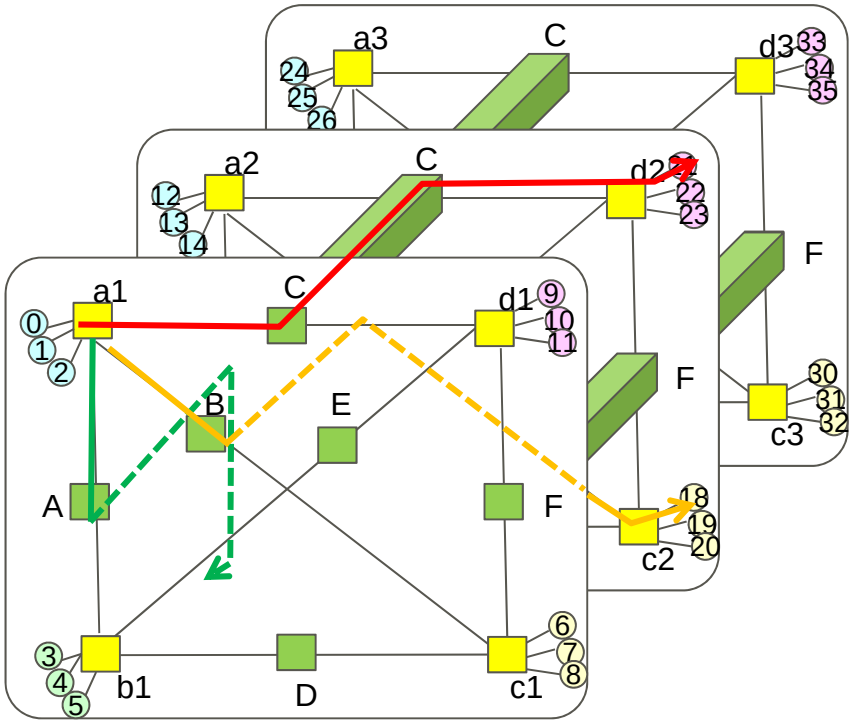


各フェーズでの送信先 \ 送信元	0	1	2	3	4	5	6	7	8	9	10	11
0	3	6	9	7	10	0	11	1	5	2	4	8
1	4	7	10	8	11	1	9	2	3	0	5	6
2	5	8	11	6	9	2	10	0	4	1	3	7
3	9	3	6	0	7	10	5	11	1	8	2	4
4	10	4	7	1	8	11	3	9	2	6	0	5
5	11	5	8	2	6	9	4	10	0	7	1	3
6	6	9	3	10	0	7	1	5	11	4	8	2
7	7	10	4	11	1	8	2	3	9	5	6	0
8	8	11	5	9	2	6	0	4	10	3	7	1
9	0	2	1	3	5	4	6	8	7	9	11	10
10	1	0	2	4	3	5	7	6	8	10	9	11
11	2	1	0	5	4	3	8	7	6	11	10	9

衝突を回避

層を跨ぐ通信

■1層目は2層目の別々の頂点に向かうようにスケジュール



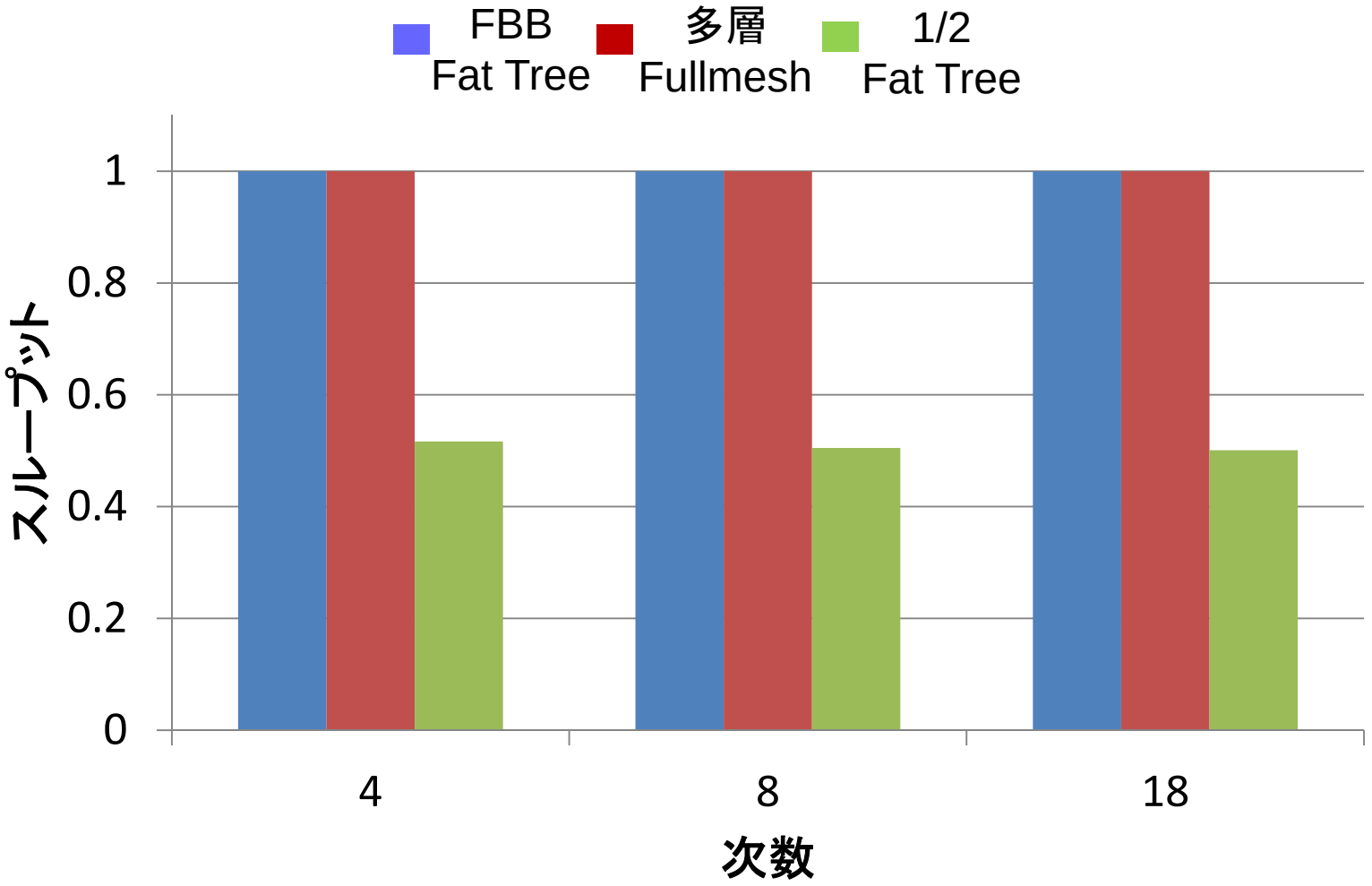
各フェーズでの送信先 \ 送信元	0	1	2	3	4	5	6	7	8	9	10	11
12	15	18	21	19	22	12	23	13	17	14	16	20
13	16	19	22	20	23	13	21	14	15	12	17	18
14	17	20	23	18	21	14	22	12	16	13	15	19
15	21	15	18	12	19	22	17	23	13	20	14	16
16	22	16	19	13	20	23	15	21	14	18	12	17
17	23	17	20	14	18	21	16	22	12	19	13	15
18	18	21	15	22	12	19	13	17	23	16	20	14
19	19	22	16	23	13	20	14	15	21	17	18	12
20	20	23	17	21	14	18	12	16	22	15	19	13
21	12	14	13	15	17	16	18	20	19	21	23	22
22	13	12	14	16	15	17	19	18	20	22	21	23
23	14	13	12	17	16	15	20	19	18	23	22	21

衝突を回避

All-to-all通信時の性能評価

FUJITSU

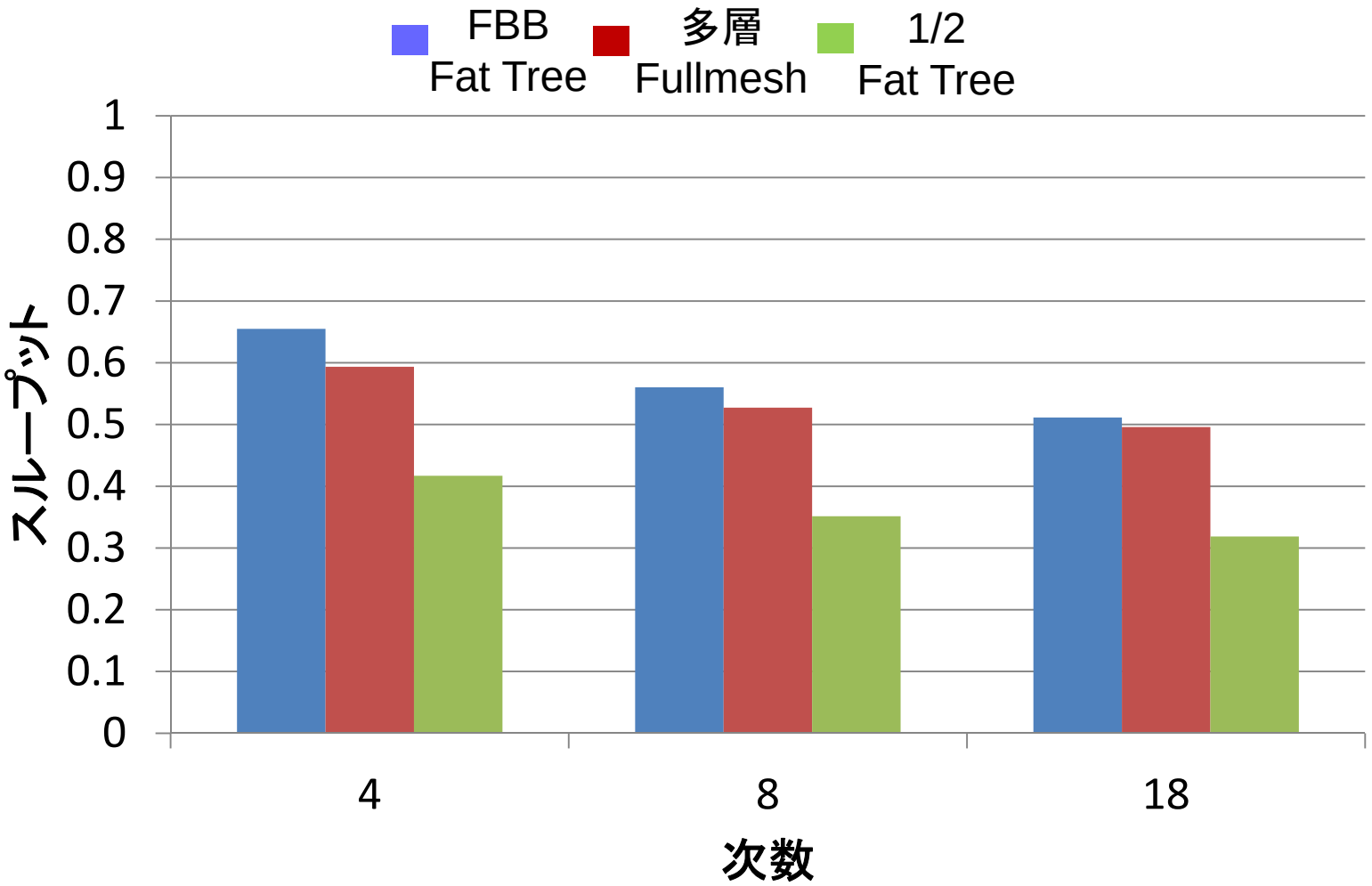
■多層FullmeshはFat Treeと同様に経路競合が回避できている



ランダム通信時の性能評価

FUJITSU

■規模が大きい場合Fat Treeと多層Fullmeshの性能は同程度



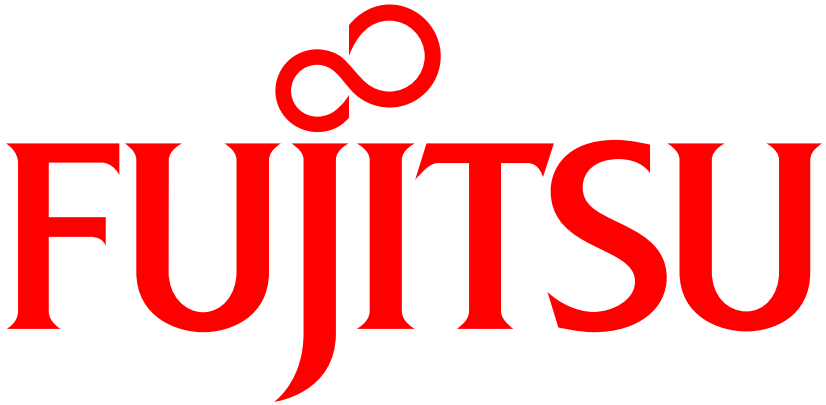
まとめと実用化に向けて

FUJITSU

- 多層Fullmeshトポロジーを提案
  - 3段Fat Treeと比較してスイッチ数を4割削減
  - All-to-all性能はFat Treeと同等

性能を維持しつつスイッチ数を削減

- 実用化に向けて
  - ジョブスケジューラとMPIのランク配置手法の確立
  - 実アプリでの効果の評価



shaping tomorrow with you