アクセラレータと
スーパーコンピュータ

天野

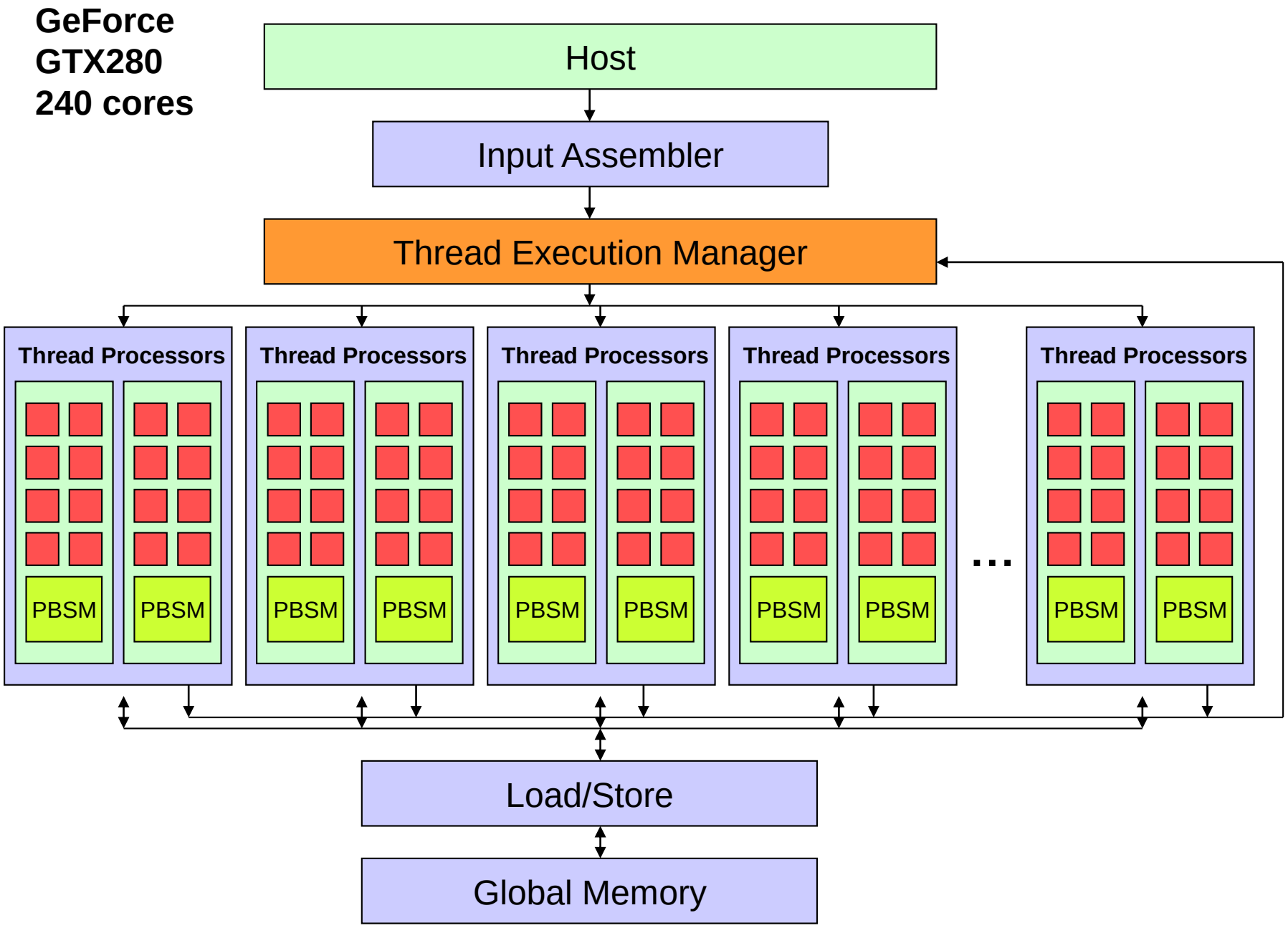
アクセラレータとは？

- 特定の性質のプログラムを高速化するプロセッサ
- 典型的なアクセラレータ
  - GPU(Graphic Processing Unit)
  - Xeon Phi
  - FPGA(Field Programmable Gate Array)
  - 最近出て来たDeep Learning用ニューロチップなどDomain Specific Architecture

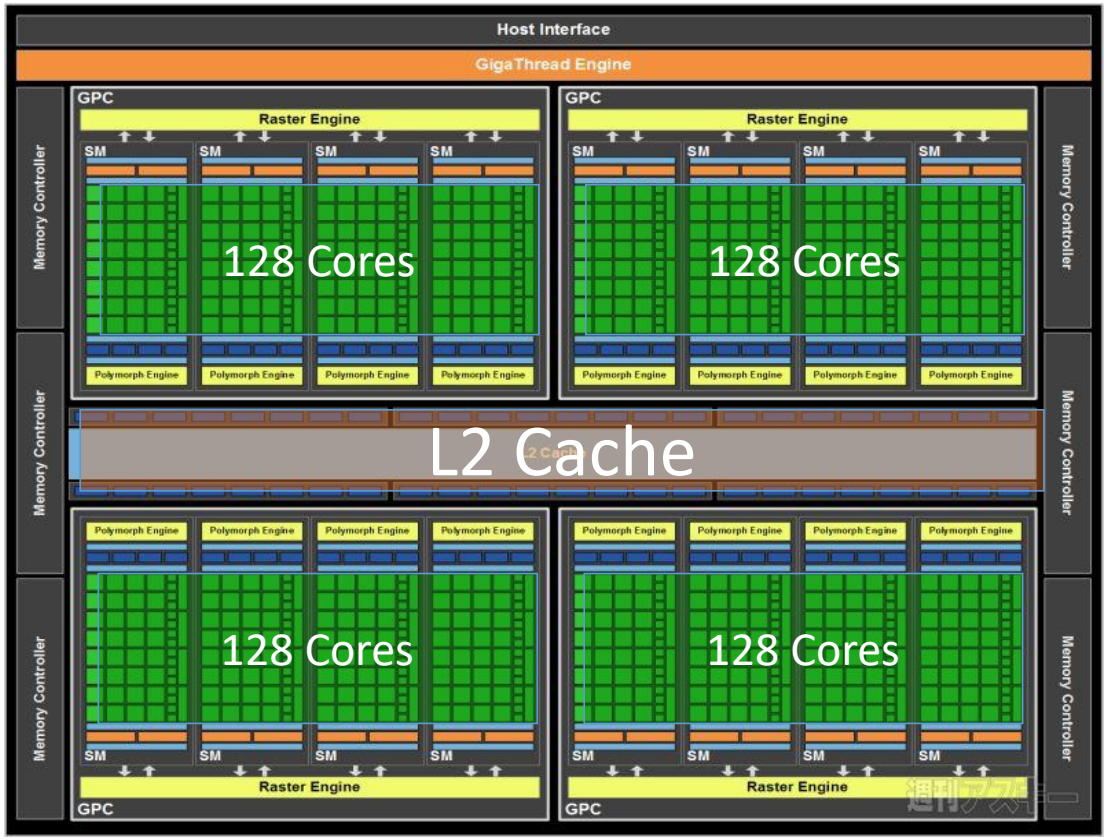
①GPGPU：General Perpose Computing with GPUグラフィックプロセッサをアクセラレータとして使う

- TSUBAME2.0(Xeon+Tesla,Top500 2010/11 4th)
- 天河一号(Xeon+FireStream,2009/11 5th)

発環境


GeForce
GTX280
240 cores
Host
Input Assembler
Thread Execution Manager
Thread Processors
Thread Processors
Thread Processors
Thread Processors
Thread Processors
PBSM
PBSM
PBSM
PBSM
PBSM
PBSM
PBSM
PBSM
PBSM
PBSM
...
Load/Store
Global Memory

GPU　(NVIDIA's GTX580)



128個のコアは
SIMD動作をする

4つのグループは
独立動作をする

もちろん、このチップを
たくさん使う

512 GPU cores (128 X 4)
768 KB L2 cache
40nm CMOS　550 mm^2

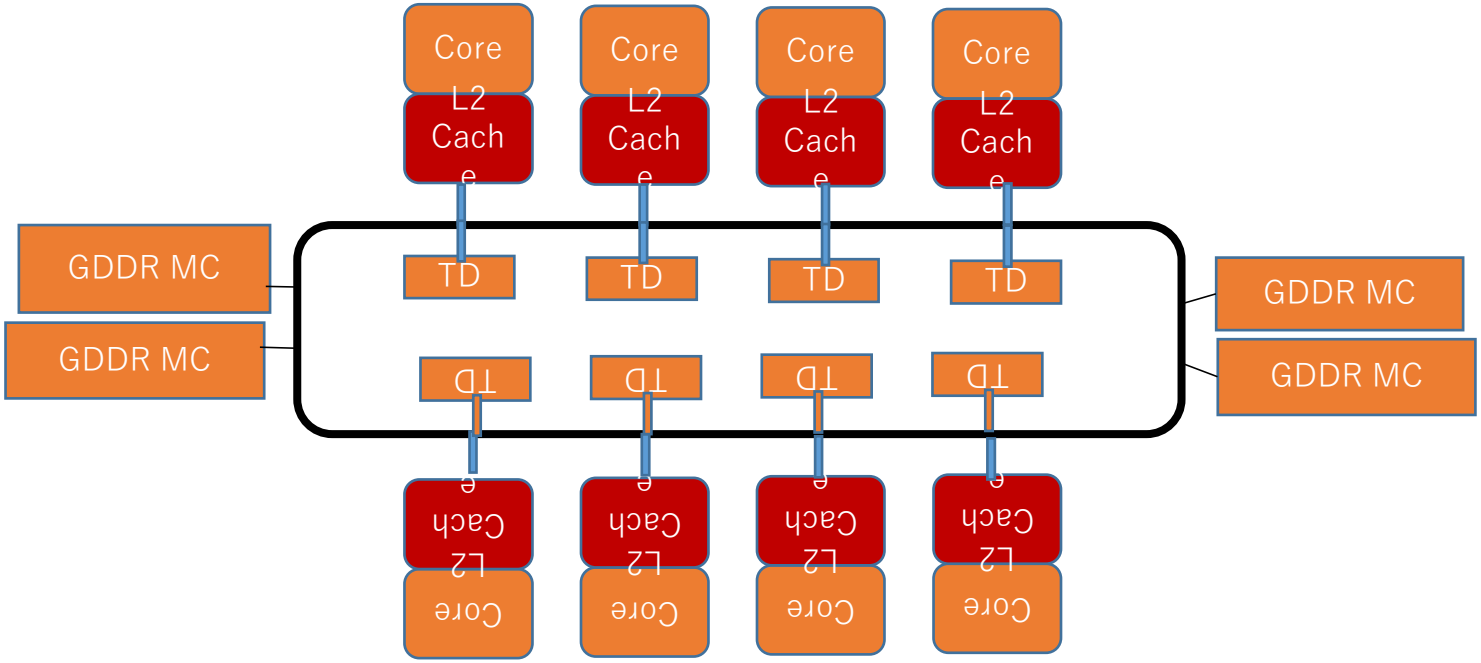
NVIDIAのGPUの名前が訳が分からん問題

- 目的用途別の名前とアーキテクチャの名前が混乱しがち
- 目的別製品シリーズの名前
  - デスクトップ用、ゲーム用：GeForce（ジーフォース）
    - GeForce GTX＞GeForce GT＞GeForceで高性能
    - TITAN Xというグラフィック用のカードがあるがこれはPascalアーキテクチャを使っている。
    - コスト性能比が高い
  - プロ用：Quadro
    - 使ったことがないので良く分からないが凄そう
  - モバイル用：Tegra
    - 車載などの用途のための低電力
    - Tegra X1:Maxwell アーキテクチャを使っている
    - Tegra K1:Keplarアーキテクチャを使っている
    - Tegra 3,2はGPUが付いていないARMだけ
  - 高性能用（AI用）：Tesla
    - 以前はGPGPU用のをTeslaと呼んでいたが最近は大きくAI用にシフトした
    - Tesla P100：Pascalアーキテクチャ
    - Tesla V100： Voltaアーキテクチャ
- アーキテクチャの名前
  - Fermi, Maxwell, Kepler, Pascal, Volta
  - プロセッサの構造を示す
  - どんどん新しいのが出てきて追従できない。。。

②NUMA型アクセラレータ：Intel Xeon Phi

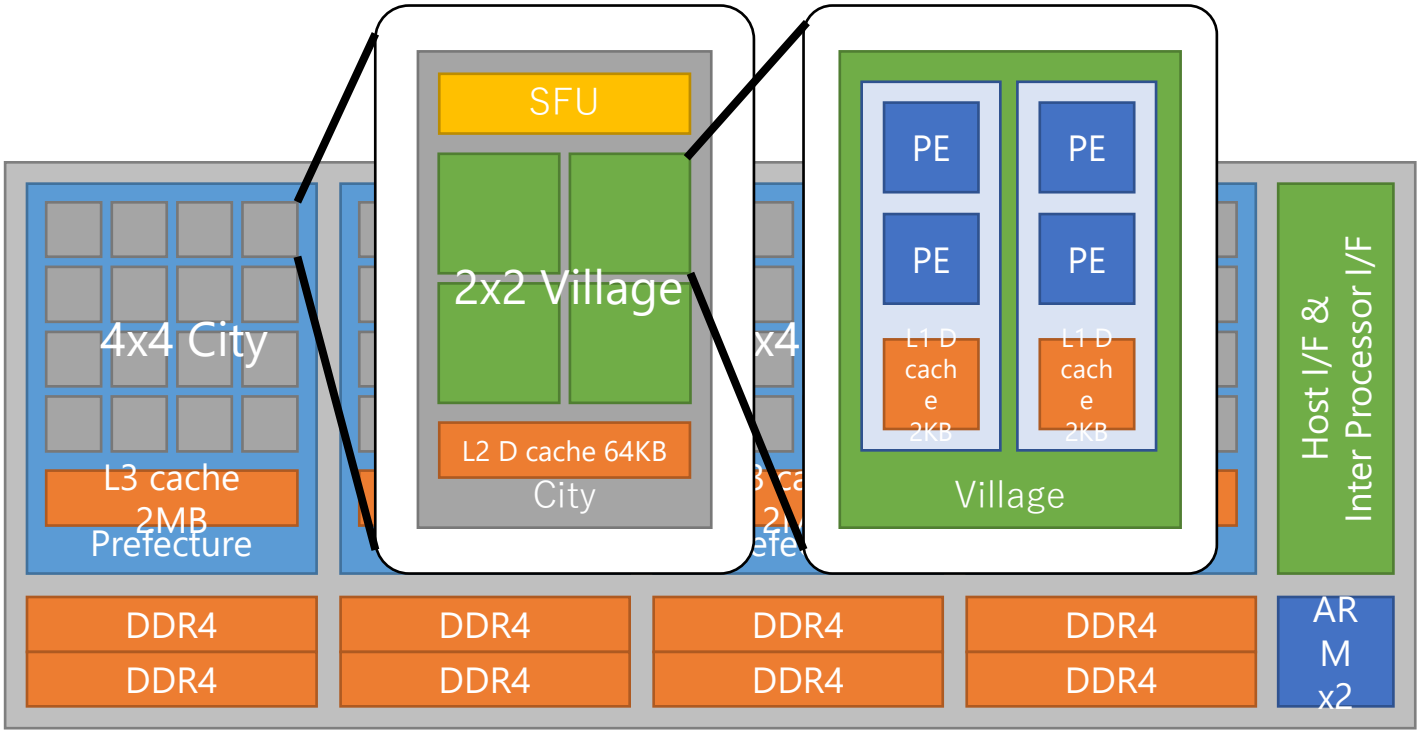
X86命令セットのアクセラレータ
一般CPUと同じコードが走る

L2キャッシュはディレクトリ方式
NUMA型



②NUMA型アクセラレータ
PEZY-SC2［Torii2017］

3-hierarchical MIMD manycore: 4PE x 4(Village) x 16(City) x 4(Prefecture)
= 1,024PE　NUMA
It provides Last Level Cache (LLC) which keeps consistency. From this viewpoint it can be CC－NUMA.



③アクセラレータとしてのFPGA

- Stratix 10
  - 14nm Intelプロセス利用
  - HyperFlexの採用によりGH z 台の動作周波数
    - 配線構造上にレジスタを置く
  - 最大10TFLOPSの浮動小数DSPモジュール
  - ARM Cortex A53 Quad Core
- Arria 10
  - 20nm TSMCプロセス
  - 最大1.5TFLOPS
  - SoCタイプはDual Core ARM Cortex

Arria10 SoCボード

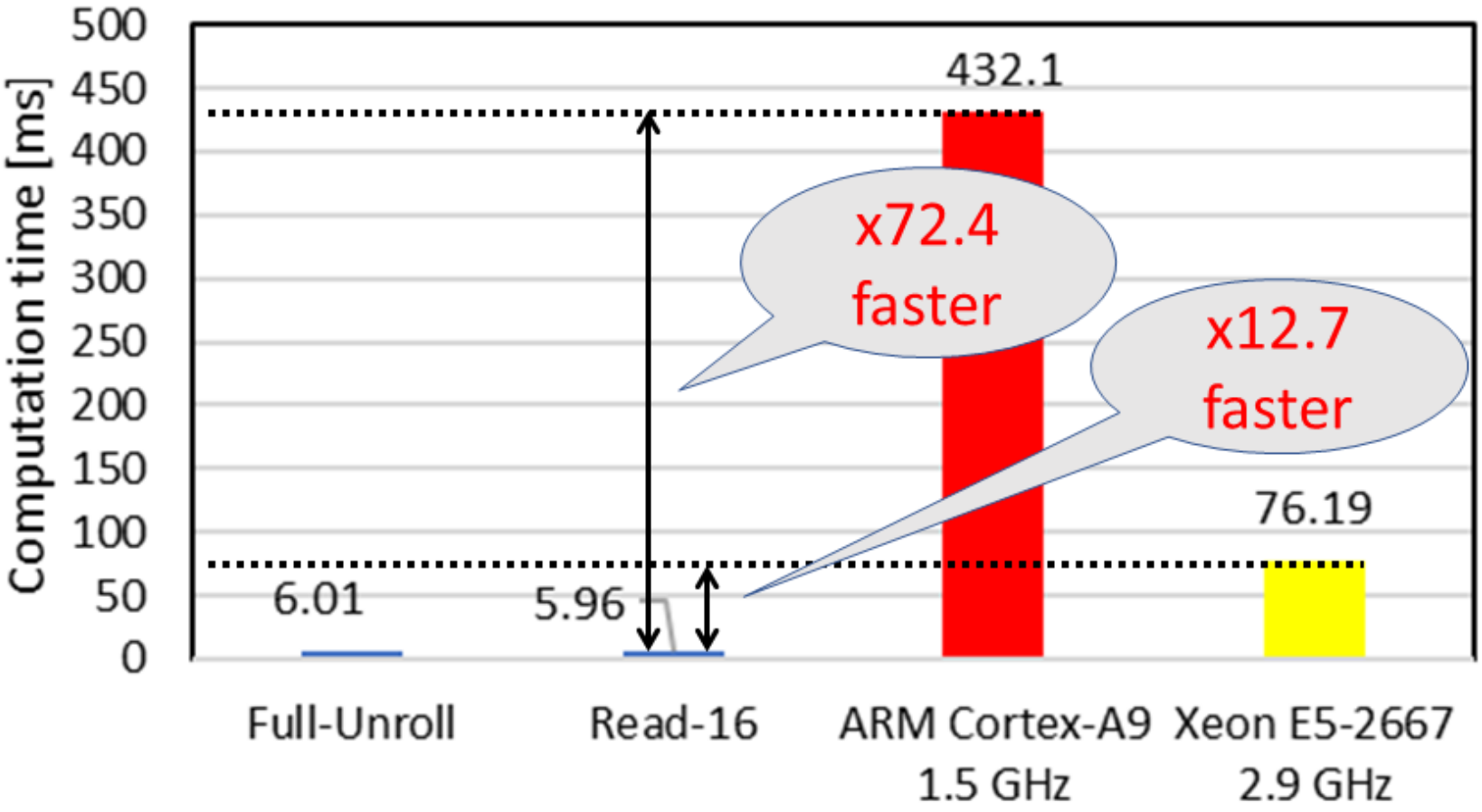
SSDより簡単にLinuxがブートする
Ethernetでネットワークに接続
内蔵ARMをホストにOpenCLでの設計ができる

ではGPUより速いのか？

	Stratix 10	Tesla P100
最大TFLOPS	9.2？	9.3
最大電力（W）	33－45	250
価格	177万（開発キット）	81万（Amazon）

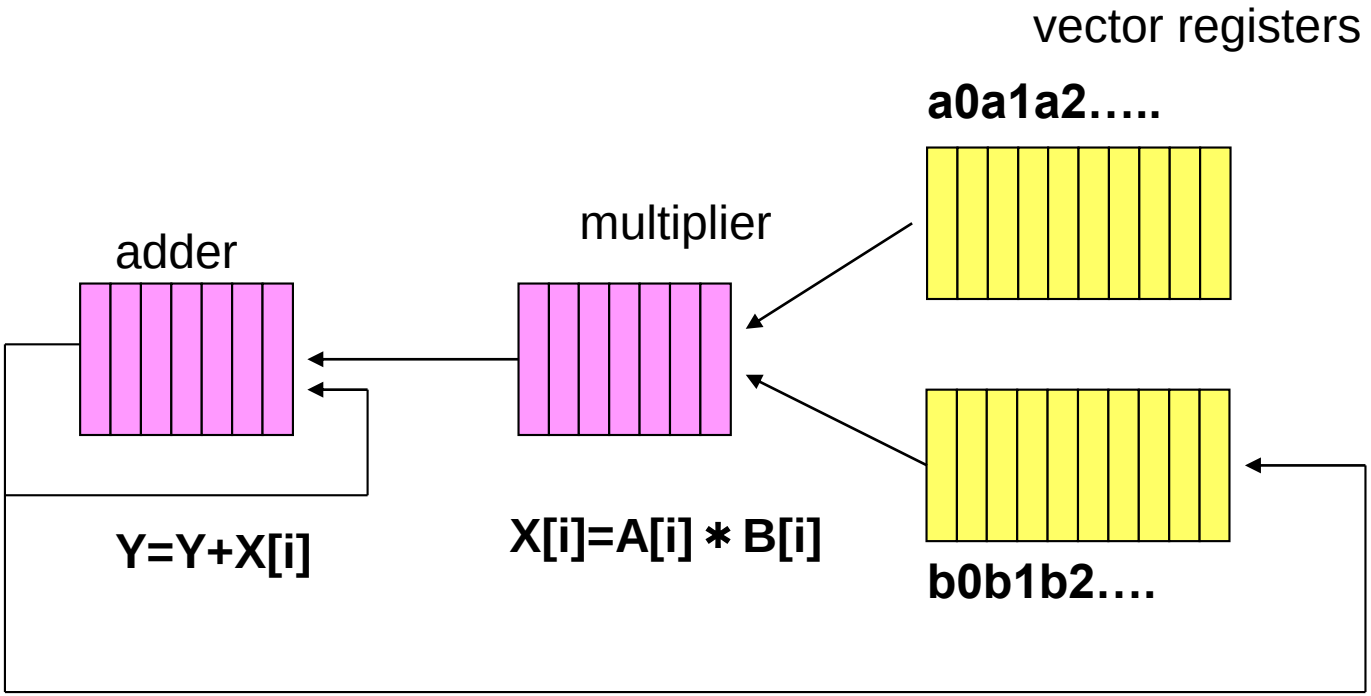
- まだStratix 10とGPUとの比較は国際会議などでは出てきていない
- Arria10の場合は、アプリケーションとチューニングのテクニックによるがGPUには絶対性能では勝てない場合が多い。しかし電力性能では勝つ
- コストは現時点では不明（出だしなので高価すぎる）だが、Stratix Vを考える（シリーズによるが1チップ当たり130万円くらいする）と、GPUに比べて倍以上するのでは？

リダクション演算の高速化



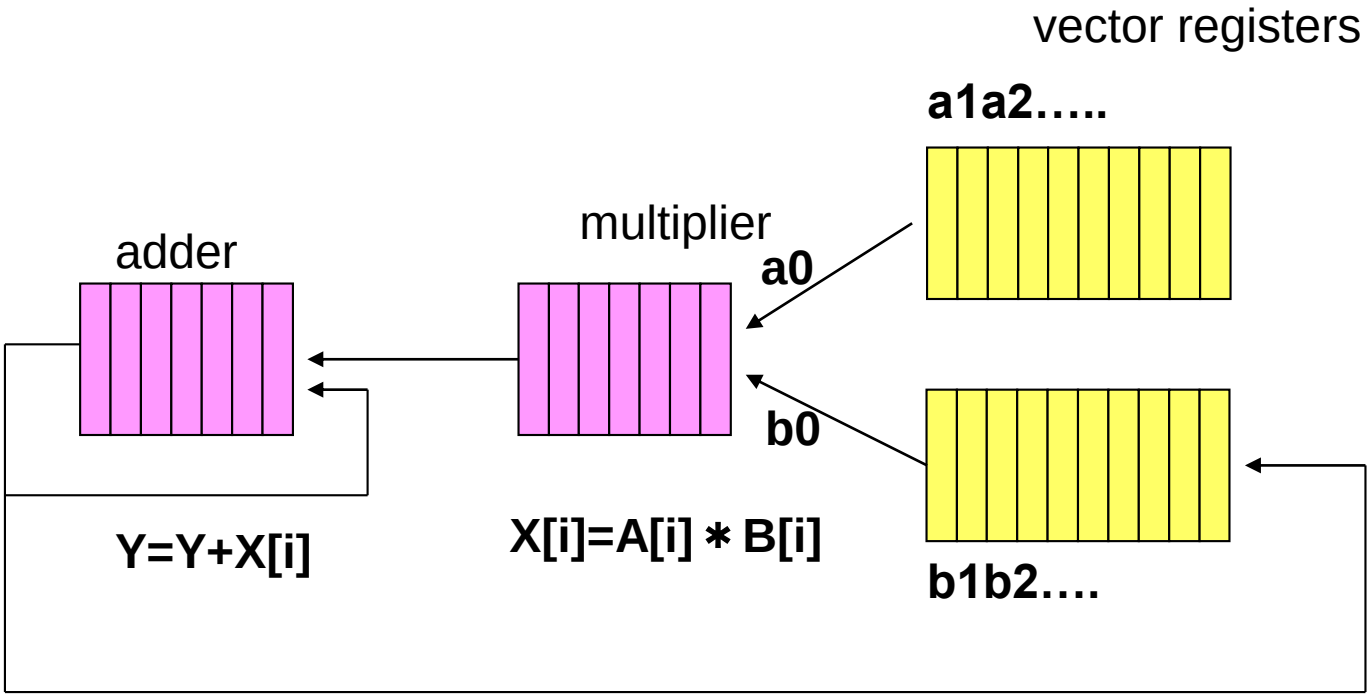
Noda Heart2017より

④ベクトル計算機

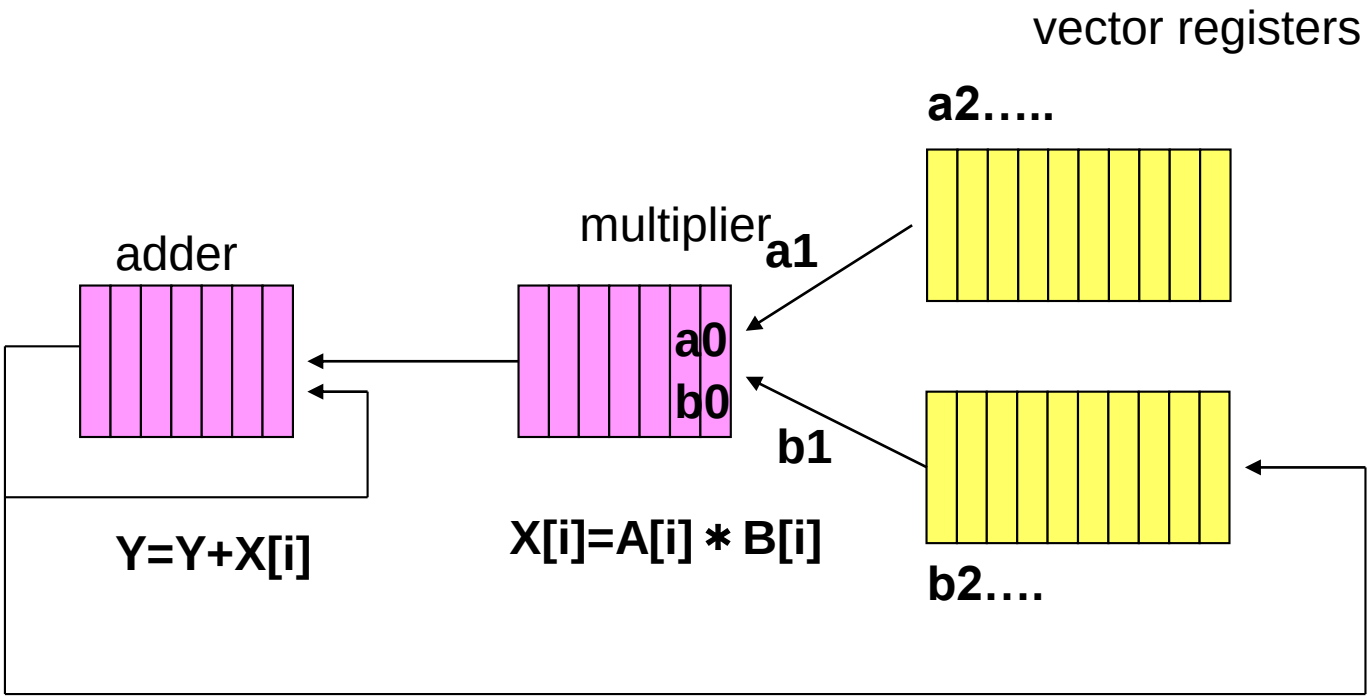


配列計算に特化した計算モジュール
Cray-1以来伝統的にスーパーコンピュータに使われる
地球シミュレータはこれを使っていた
今でもNECが得意としている→アクセラレータAurora

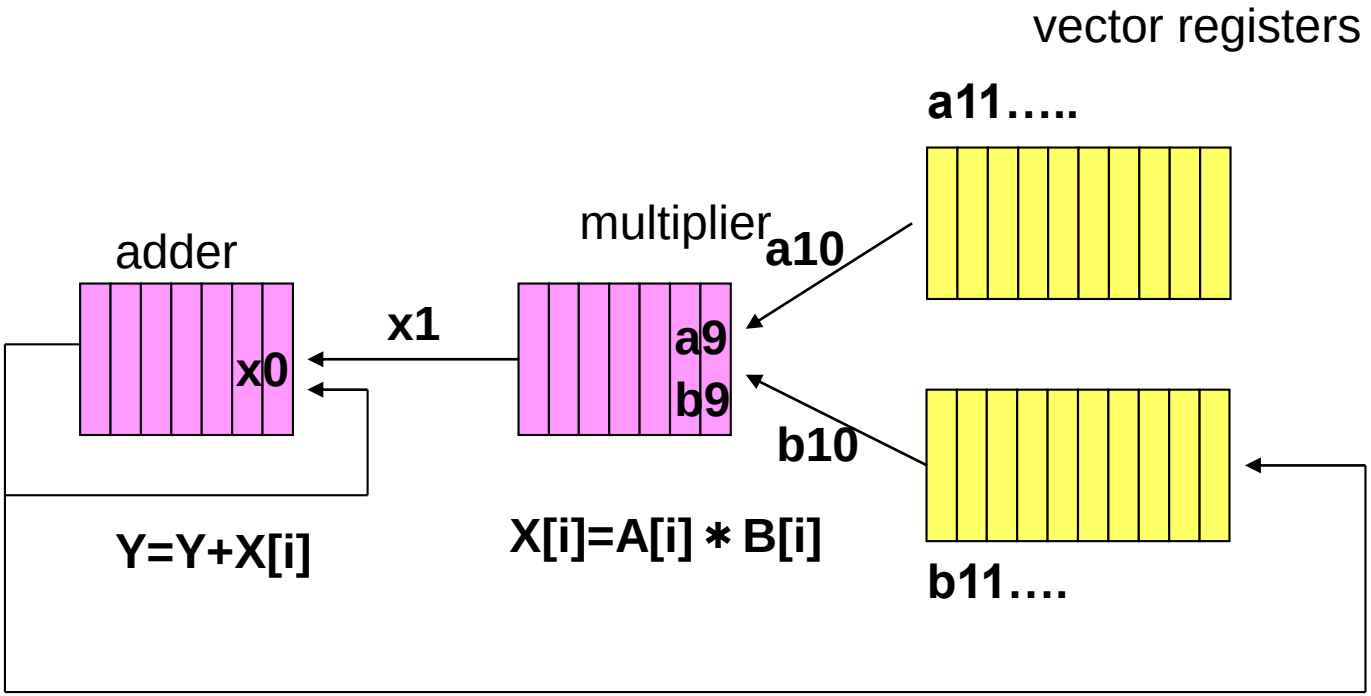
ベクトル計算機


vector registers
a1a2…..
multiplier
adder
a0
b0
Y=Y+X[i]
X[i]=A[i] * B[i]
b1b2….

ベクトル計算機

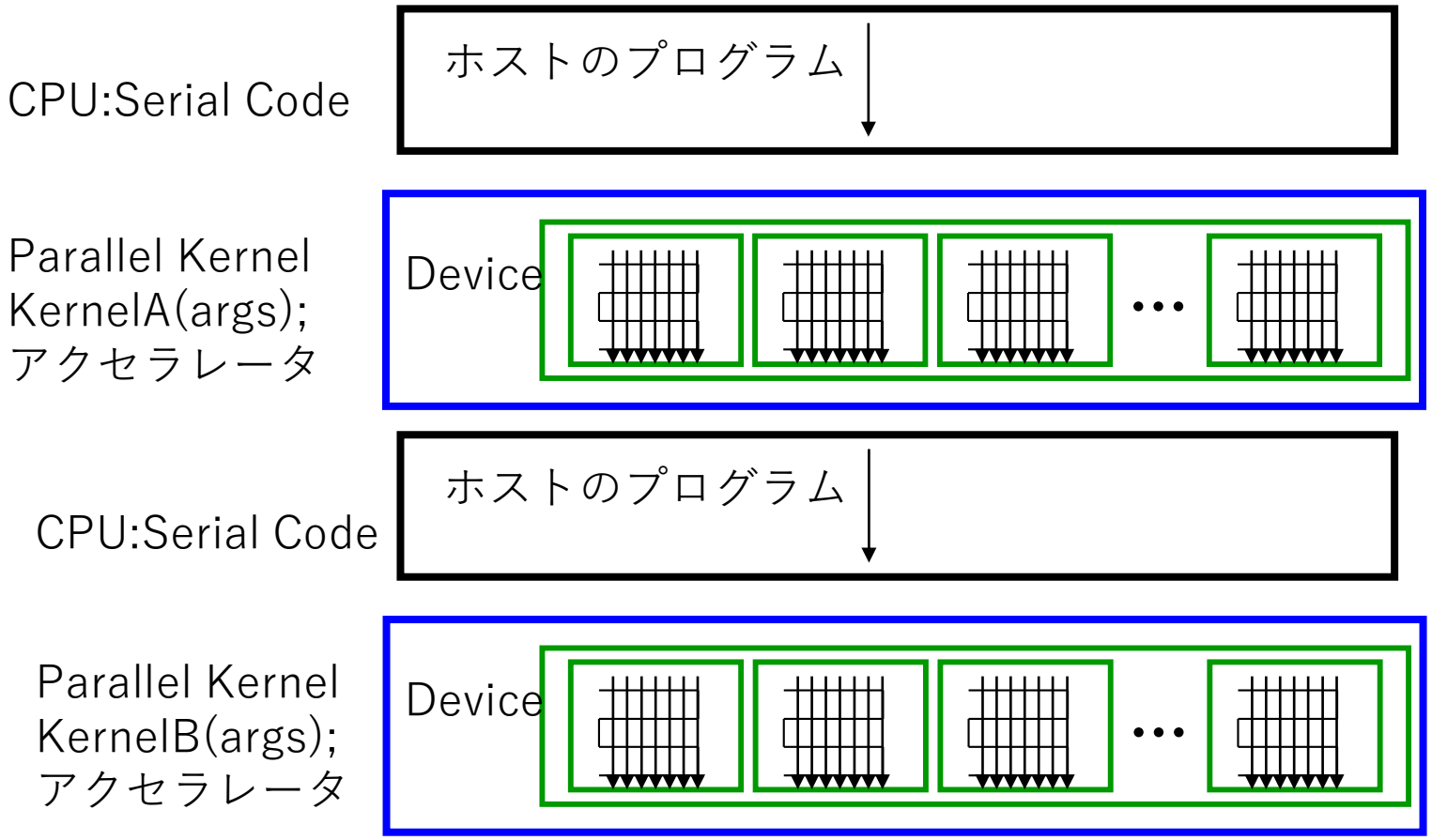

vector registers
a2…..
multiplier
a1
adder
a0
b0
b1
Y=Y+X[i]
X[i]=A[i] * B[i]
b2….

ベクトル計算機



メモリのバンド幅をGPUほど要求しない
高い平均性能が得られる
一定数のファンが居る

アクセラレータのプログラム



ホストのプログラムが準備してアクセラレータのプログラムにデータを渡す
処理が終わったら回収
CUDA（後で演習をやる）、OpenCLはこの考え方を取る

スーパーコンピュータとは？

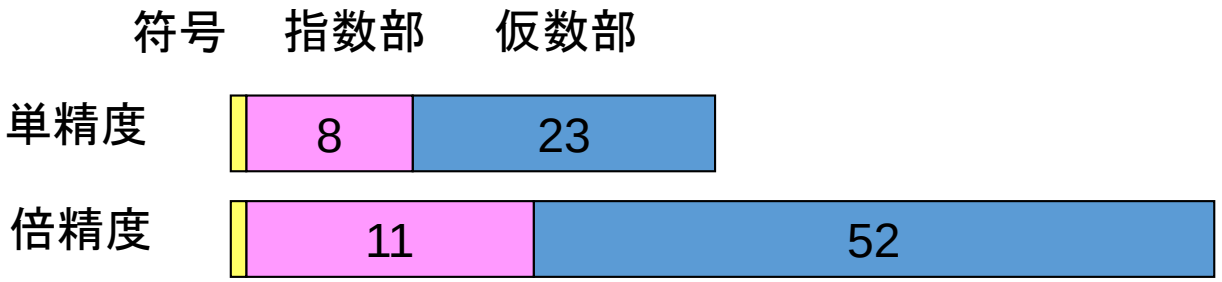
- 科学技術計算用の高性能コンピュータ
  - 物理学、天文学、気象学、地震、海洋、創薬そのほか様膨大な計算能力が必要なアプリケーションを実行
  - 非常に高価、国のお金が必要
  - 開発と管理には高い技術が必要
  - 中国、米国、日本がトップ１を争う。ヨーロッパがTop10に数台ランクイン
  - 多額の国家予算が投入されるため、一般の関心が高い。

例）

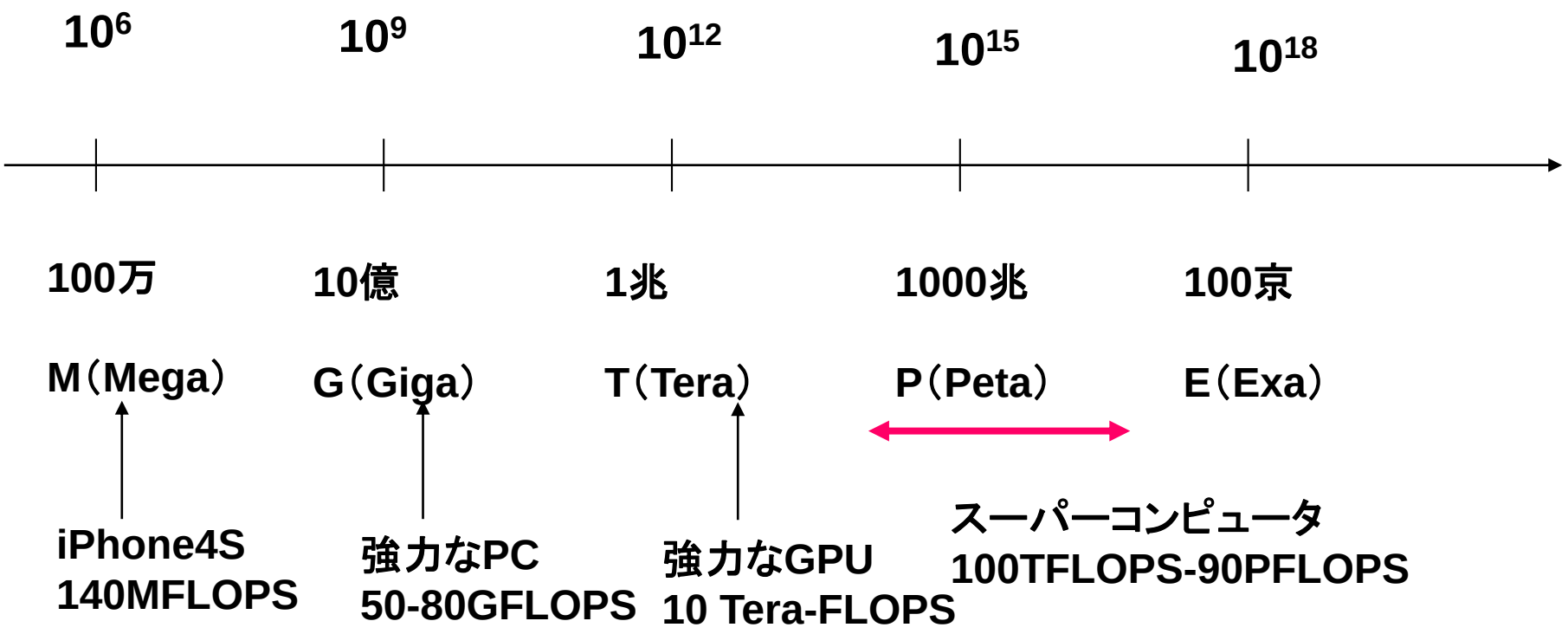
- スーパーコンピュータ京の事業仕分け（2009年）
- Pezyスパコンの研究費詐欺事件（2017年）

FLOPSとは？

- 1秒に何回浮動小数点演算ができるか？（Floating Point Operation Per Second）
- 浮動小数点数とは？
  - 仮数　×　2（指数）
  - IEEE標準でフォーマットと丸めが定義されている
  - 倍精度：64bit, 単精度： 32bit.



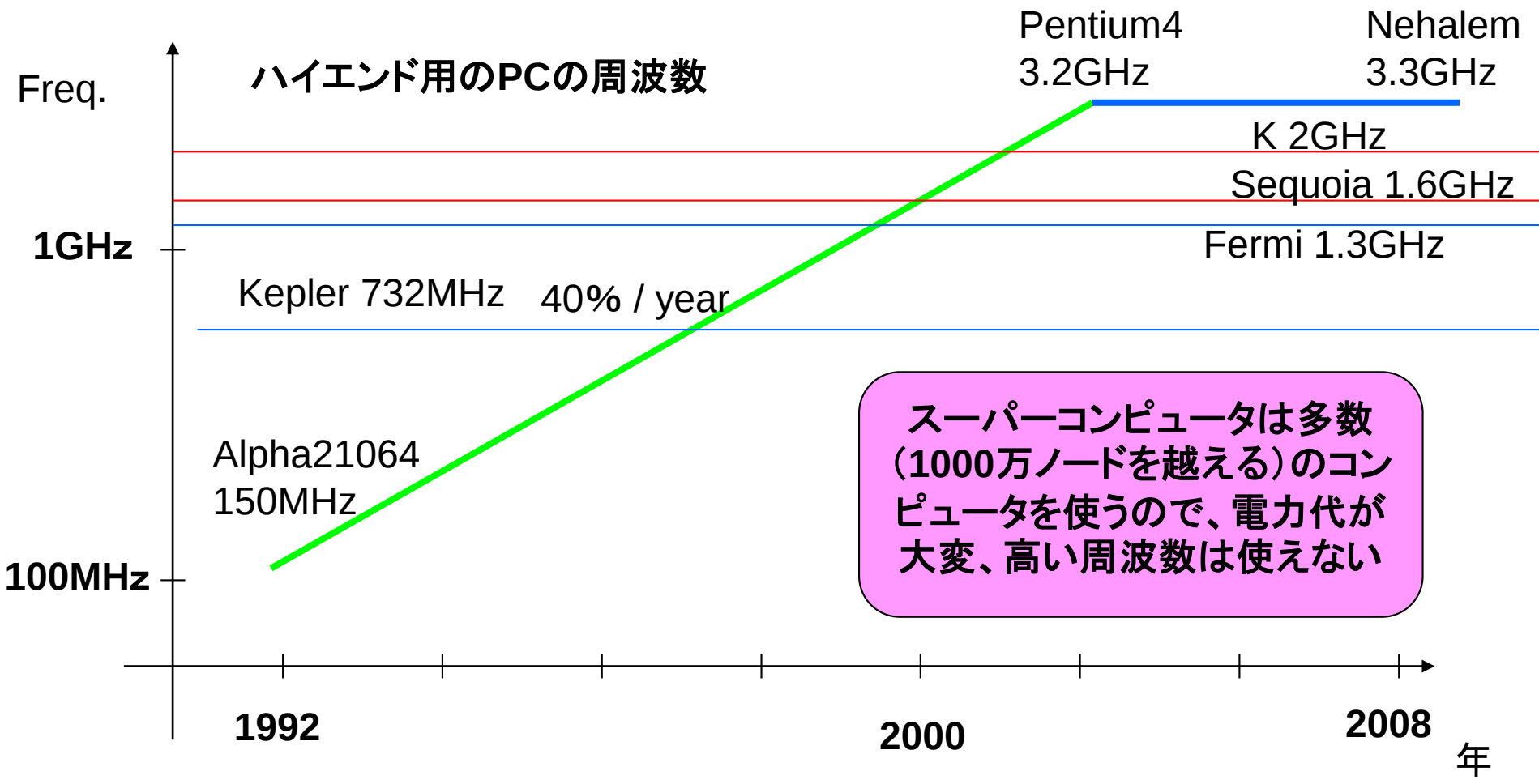
FLOPSで見るスパコンの性能



10PFLOPS ＝ 1京回
→「京」の名前の由来.

スーパーコンピュータが速いのは クロック周波数が高いせい？

× これは嘘で、普通のPCの方が高い



スパコンはなぜ速いのか？

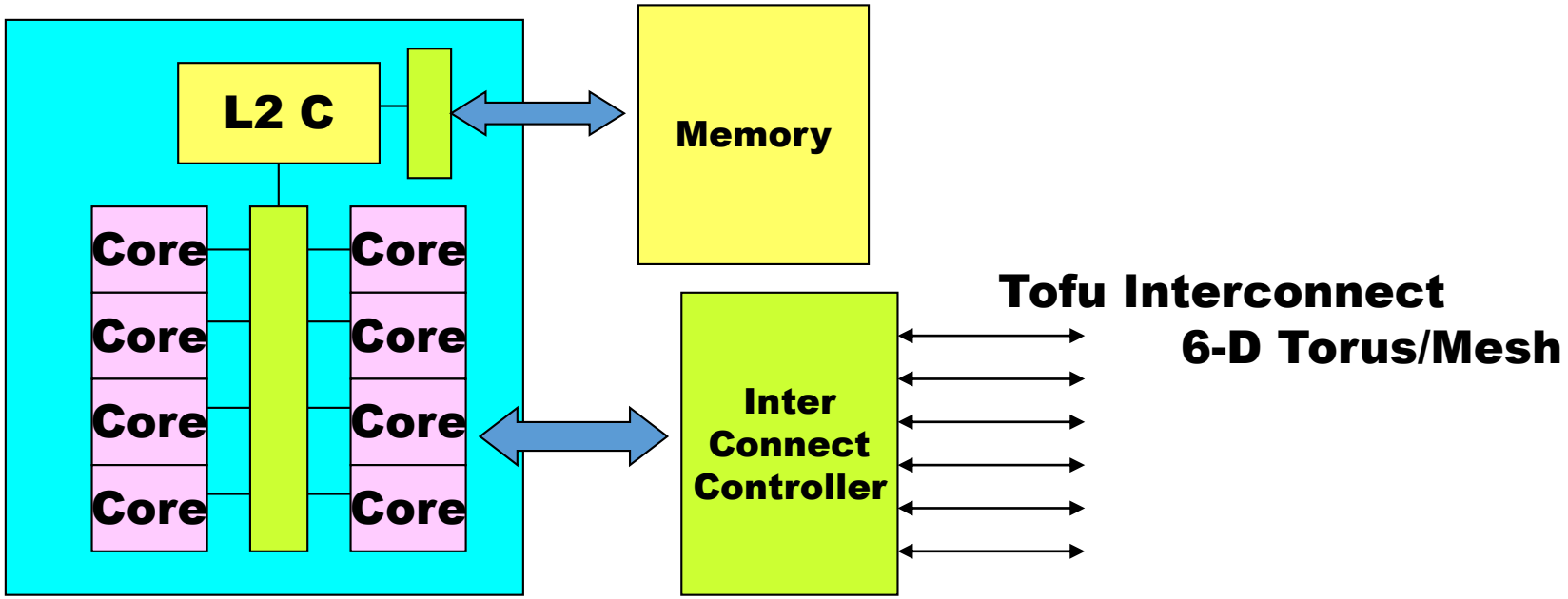
- ノード数を非常に多数持っているから
  - 世界一の性能を持つTaihuLightは1000万個もっている
- スパコンで重要なこと
  - 性能にバランスしたメモリ、接続網を持つこと
  - 十分な信頼性を持つこと
  - 安定して動作するための冷却装置、電源装置を持つこと
  - 性能を引き出すソフトウェアが整備されていること
  - 膨大な開発コストに見合うだけの価値のある利用をするための体制ができていること
  - 管理するためのインフラ的な経費と人的経費が維持可能なものであること

  →一部はデータセンターと同じだが、基本的にスパコンは商業的に採算を取るのが極めて難しい

ホモジーニアス型　対　アクセラレータ型

- ホモジーニアス型
  - スーパーコンピュータのマルチコアCPUチップを開発
  - 均質なノードを多数に接続する
  - 様々な問題で高い性能が実現できる
  - プログラミングが比較的簡単
  - 高次元トーラスが結合網として良く用いられる
  - 開発コストが膨大になる
  - TaihuLight、Sequia(IBM BlueGeneQ、Kなど
- ヘテロジーニアス型
  - アクセラレータを利用する
  - ツボにはまれば高い性能が得られるが、一般的には最大性能と実効性能の差が大きい
  - 汎用に使われているアクセラレータを利用するので開発コストが小さい
  - Infiniband＋Fat Tree、Dragonflyなどが結合網として利用される
  - 性能対電力効率が優れている場合が多い。
  - Tenhe-2(Xeon Phi), Titan(Kepler), Gyoukou(Pezy-SC2)

ホモジーニアス型の例「K」



SPARC64 VIIIfx Chip

4 nodes/board

24boards/Lack

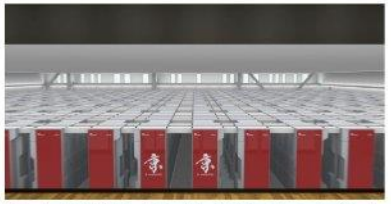
96nodes/Lack

RDMA mechanism
NUMA or UMA+NORMA

「京」のシステム構成

システム全体

計算ラック×864



1京回/秒≒10.6ペタフロップス
1PB以上

計算ラック群

計算ラック×8
ディスクラック×2

98.4兆回/秒
12TB

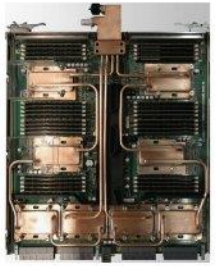
計算ラック

システムボード×24
IOシステムボード×6

12.3兆回/秒
1.5TB

システムボード

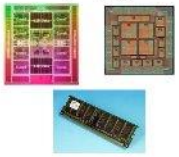
ノード×4



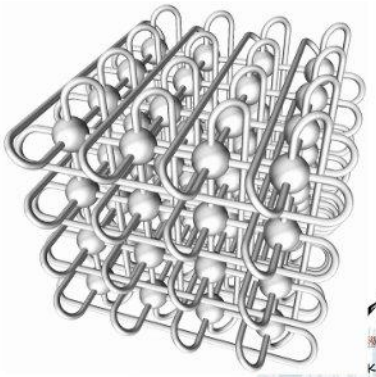
5120億回/秒
64GB

ノード

CPU×1
ICC×1
メモリ



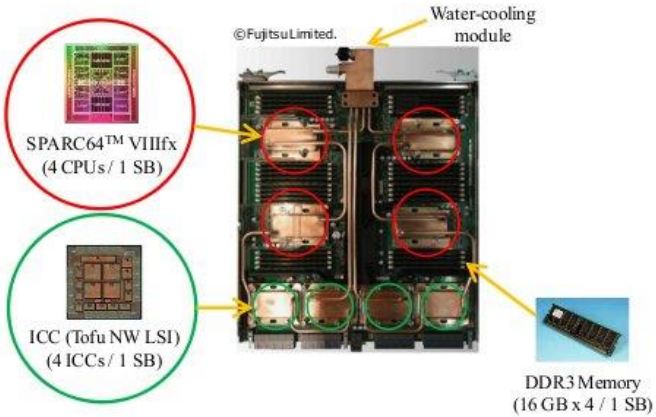
演算性能:1280億回/秒
メモリ容量:16GB



SACSIS2012 Invited speech

CPUの詳細

	諸元
演算性能 （ピーク）	128 GFLOPS (16 GFLOPS x 8 cores)
コア数	8
クロック周波数	2.0 GHz
浮動小数点 演算器	乗加算ユニット x 4 (2 SIMD) 除算器 x 2
レジスタ数	浮動小数点レジスタ (64bit) : 256 汎用レジスタ (64bit) : 188
キャッシュ	L1I$: 32 KB (2way) L1D$: 32 KB (2way) L2$: Shared 6 MB (12way)
メモリ帯域	64 GB/s (0.5B/F)

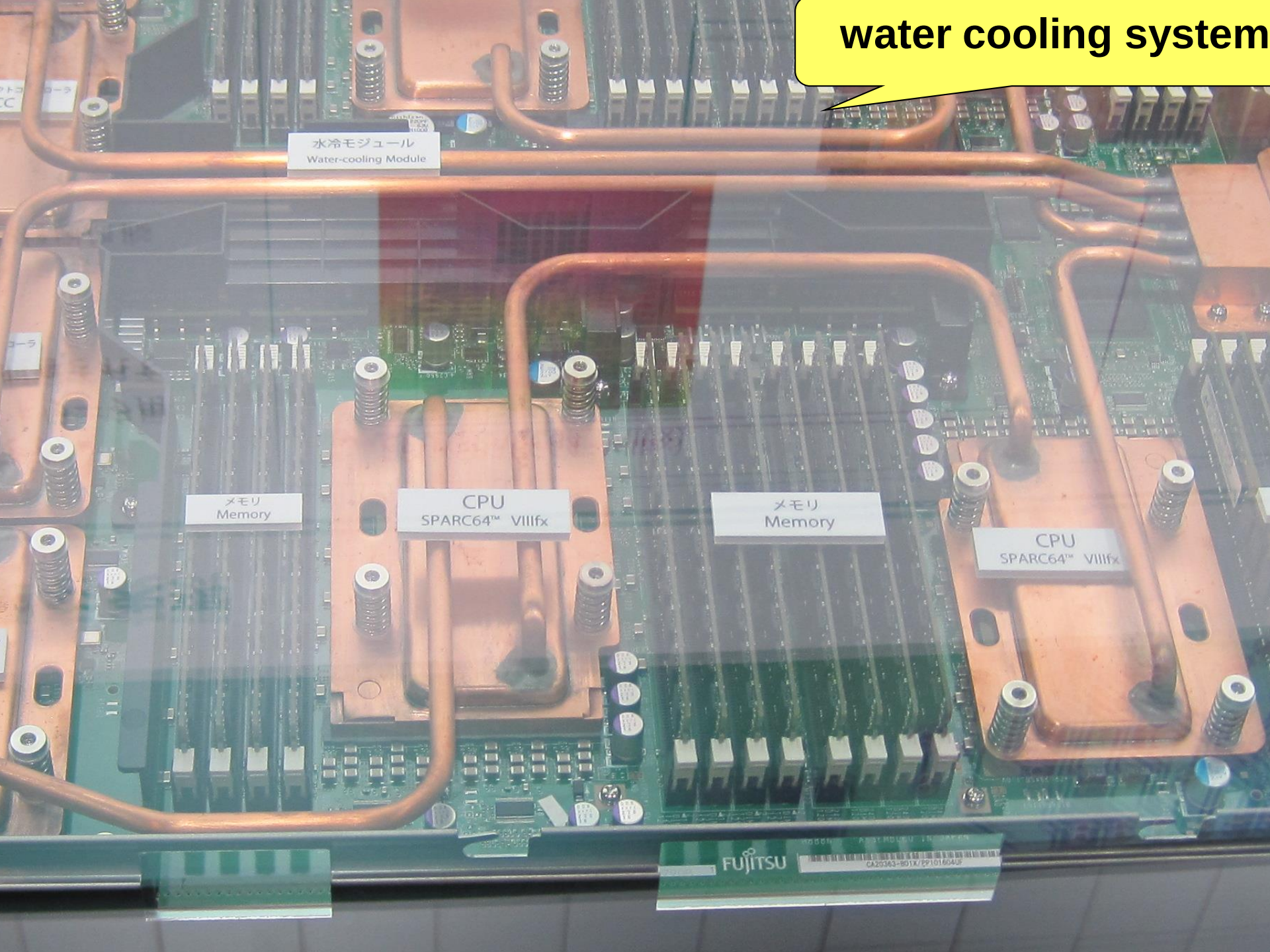


45nm CMOS process
チップサイズ: 22.7mm x 22.6mm
トランジスタ数 : 760M
Power : 58W (TYP, 駆動温度30℃), 水冷

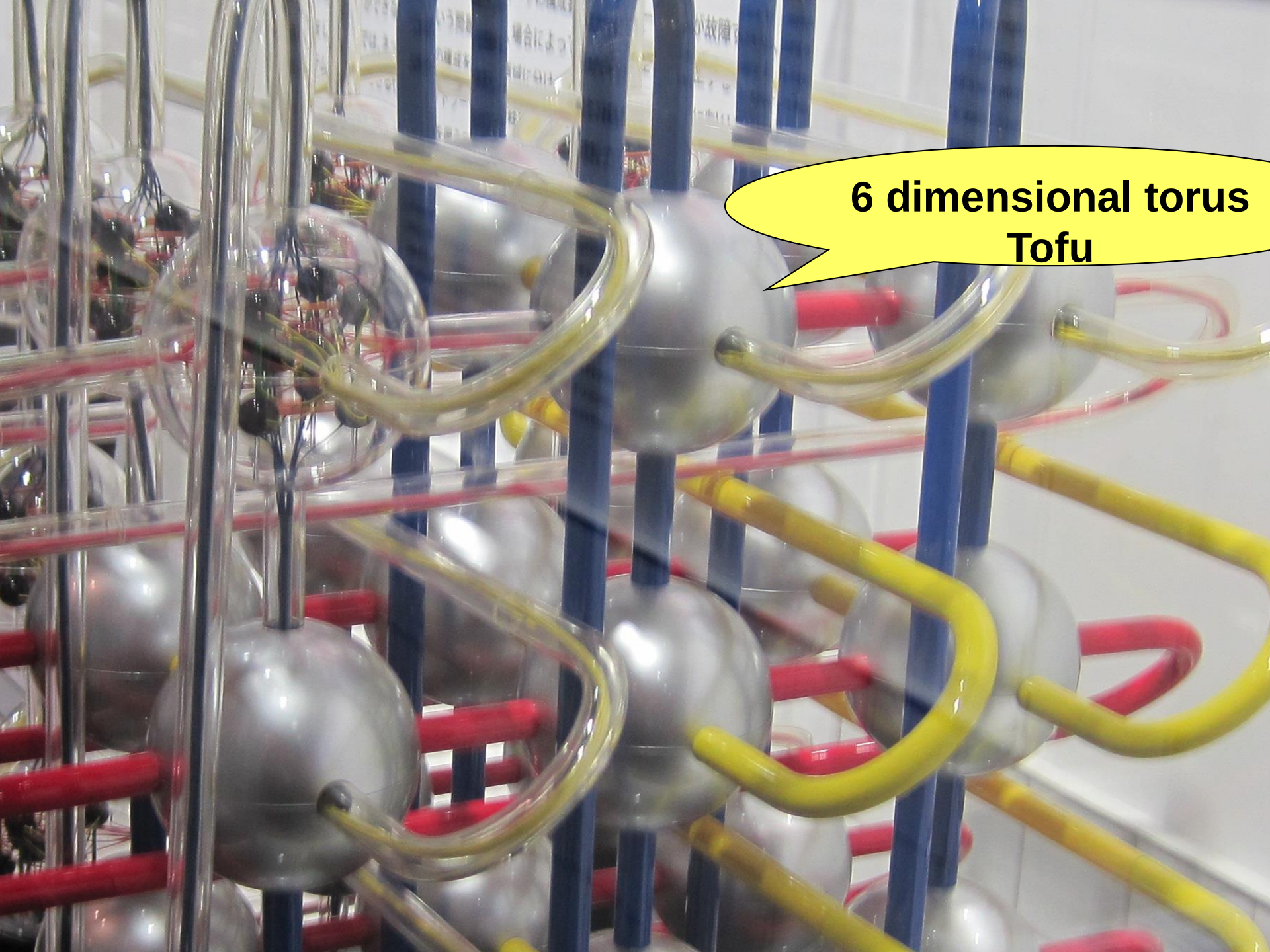
他のチップとの比較

Vendor	Name	Core	Process rule (nm)	Peak performance (GFLOPS)	Cache (MB)	Power (W)	GF/W	System (w/planned)
IBM	PowerPC A2	16	45	204.80	32	55	3.72	Sequoia (BlueGene/Q)
Intel	E3-1260L	4	32	105.60	8	45	2.35	
Fujitsu	SPARC64VIIIfx	8	45	128.00	6	58	2.21	K computer
IBM	Power7	8	45	256.00	32	200	1.28	
AMD	Opteron 6172	12	45	100.80	12	80	1.26	XE6,etc.
Intel	Xeon X5670	6	32	79.92	12	95	0.84	TSUBAME2.0,etc.

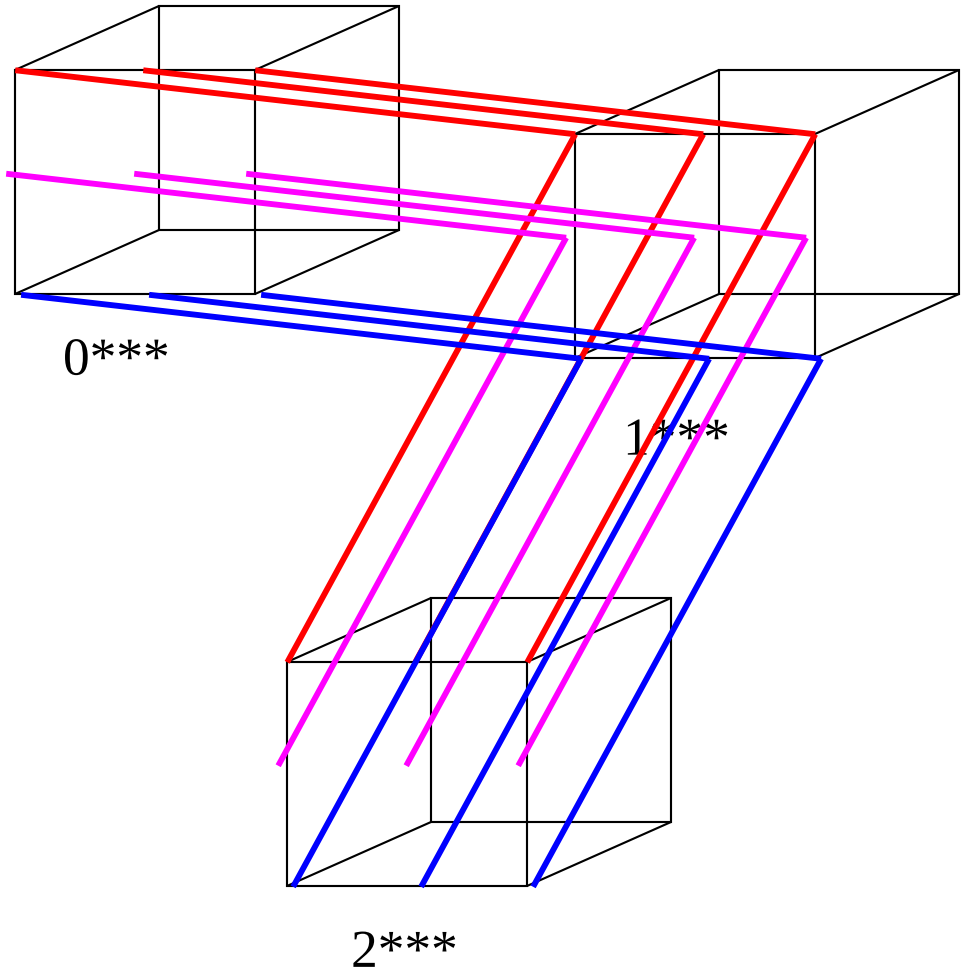
高性能かつ低消費電力


water cooling system
水冷モジュール
Water-cooling Module
メモリ
Memory
CPU
SPARC64™ VIIIfx
メモリ
Memory
CPU
SPARC64™ VIIIfx
FUJITSU

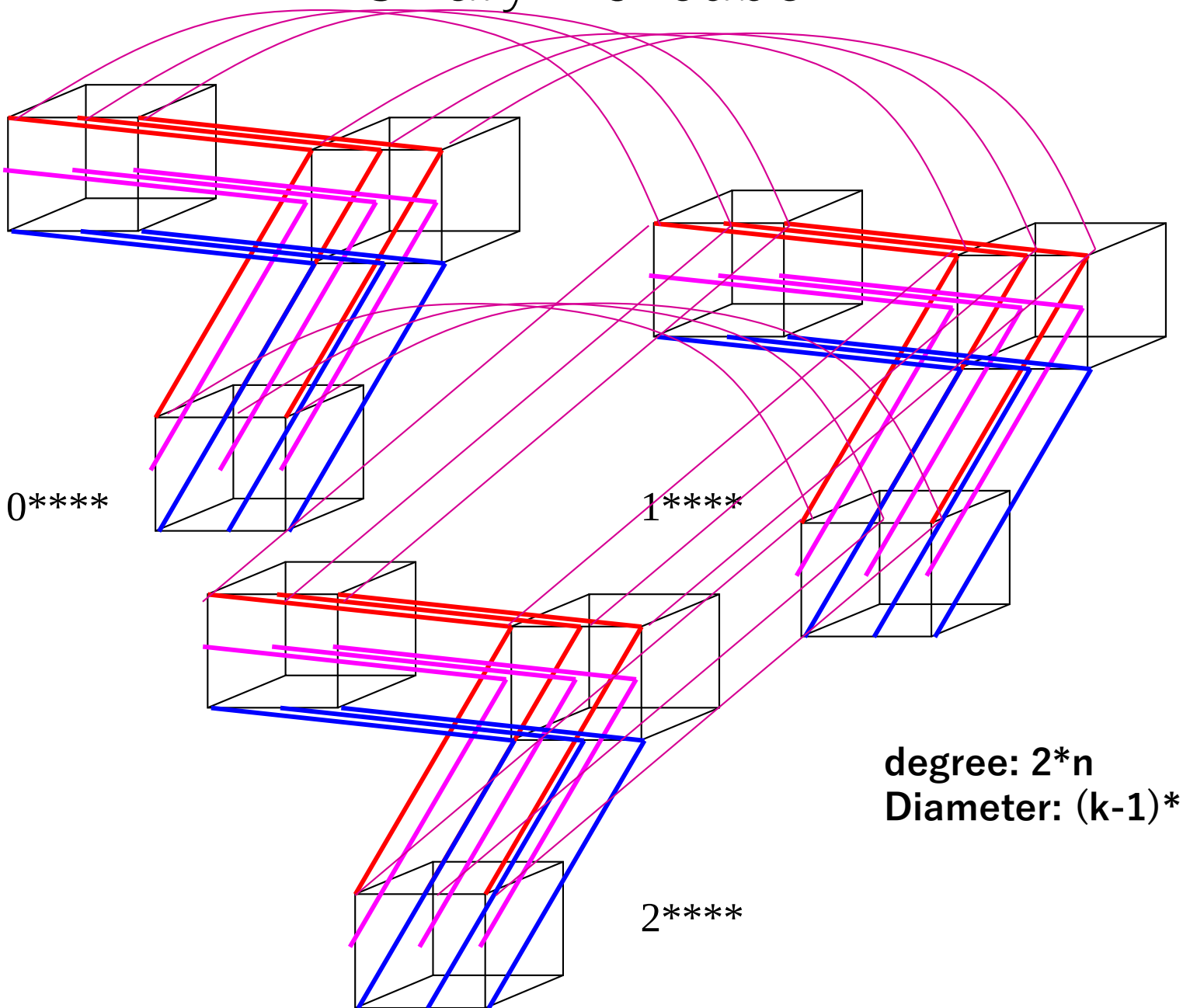
Lacks of K


6 dimensional torus
Tofu

3 -ary 4-cube



3 -ary 5-cube

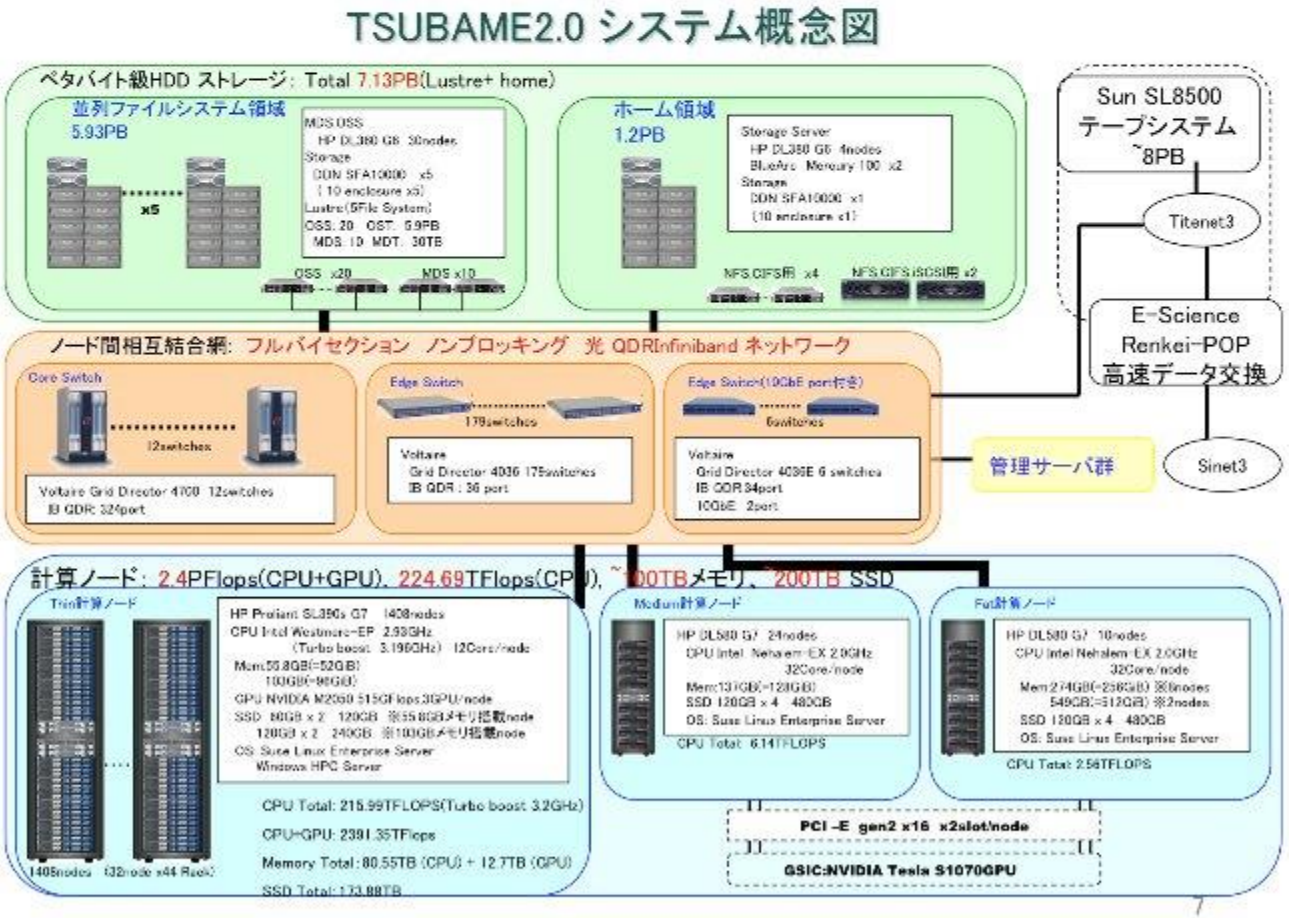


degree: 2*n
Diameter: (k-1)*n

TSUBAME2ハードウェア構成

東京工業大学
Tsubame GPUsを用いた
ヘテロジーニアス型

TSUBAME2ではおよそ1400台の計算ノードがフルバイセクションバンド幅の
計算ノードはThin, Medium, Fatの3種類用意されており、それぞれ搭載メモリ
よそ7ペタバイトの大容量ストレージシステムが接続されており、並列ファイルシステム領域とホーム領域から構成されます。



どうやってTop1を決めるのか？

- **Top500/Green500: Linpackという行列計算のカーネルを実行して性能／電力当たりの性能を測定する。**
  - **Weak Scaling:対象の行列のサイズはいくらでも大きくできる**
  - **演算性能重視のアプリケーション、メモリ、転送性能はあまり効かない**
  - **Green500はTop500に性能がランクインしていなければならない**
- **Gordon Bell Prize**
  - **ピーク性能、性能価格比、特殊な性能などユニークな特徴を表彰**
- **HPC Challenge**
  - **Global HPL　行列計算: 演算性能重視**
  - **Global Random Access：ランダムメモリアクセス：通信性能重視**
  - **EP stream per system:メモリアクセス：メモリ性能重視**
  - **Global FFT: 複雑な問題で通信とメモリの両方を測定**

- **11月　ACM/IEEE Supercomputing Conference**
  - **Top500、Gordon Bell Prize、HPC　Challenge、Green500**
  - **今年はDallas**
- **6月　International Supercomputing Conference**
  - **Top500、Green500**
  - **今年はFrunkfurt**

Top500の変遷

- 2014年にTenhe-2が登場して以来はトップは中国
  - 2位に倍程度の性能差を付けており、TaihuLightが登場しなければ今でも首位
  - 中身はIntelのXeon-Phiのヘテロジーニアス型
- 2016年にTaihuLightが登場
  - 中国オリジナルの260コアを持つ ShenWei (神威) を利用のホモジーニアス型
- 青地（ホモジーニアス専用型）と赤字（アクセラレータ利用型）が混在
- 最近は上位5位は安定し、それ以降には大きな変化がある
  - どうも20PFLOPS位に壁があるようで現在の所中国しか越えていない
- Green500は、かつてはGPUを用いたものが独占していたが、Pezy-SCが上位を占めるようになった
  - Green500は評価が難しい。。。

Top 500 2015 July

Name	Development	Hardware	Cores	Performance TFLOPS	Power（KW)
Tianhe-2(天河) (China)	National University of Defence Technology	Intel Xeon E5-2692 12C 2.2GHz,TH Express-2, Intel Xeon Phi31S1P	3120000	33862.7 (54902.4)	17808
Titan (USA)	DOE/SC/Oak Ridge National Lab.	Cray XK7, Opteron 6274 16C 2.2GHz, Cray Gemini Intercon.NVIDIA K20x	550640	17590 (27112.5)	8209
Sequoia (USA)	DOE/NNSA/LLNL	BlueGene/Q,Power BQC 16C 1.6GHz	1572864	17173.2 (20132.7)	7890
K （京) (Japan)	RIKEN AICS	SPARC VIIIfx 2.0GHz Tofu Interconnect Fujitsu	705024	10510 (11280)	12659.9
Mira (USA)	DOE/SC/Argonne National Lab.	BlueGene/Q Power BQC 1.6GHz	786432	8586.6 (10066.3	3945

Top 5 は2013から2015年まで同じだった

Top 500 2016 July

Name	Development	Hardware	Cores	Performance TFLOPS	Power（KW)
TaihuLight(太湖之光)	National Supercomputing Center in Wuxi	ShinWei(神威) NRCPC	10649600	93014.6	15371
Tianhe-2(天河) (China)	National University of Defence Technology	Intel Xeon E5-2692 12C 2.2GHz,TH Express-2, Intel Xeon Phi31S1P	3120000	33862.7 (54902.4)	17808
Titan (USA)	DOE/SC/Oak Ridge National Lab.	Cray XK7, Opteron 6274 16C 2.2GHz, Cray Gemini Intercon.NVIDIA K20x	550640	17590 (27112.5)	8209
Sequoia (USA)	DOE/NNSA/LLNL	BlueGene/Q,Power BQC 16C 1.6GHz	1572864	17173.2 (20132.7)	7890
K（京) (Japan)	RIKEN AICS	SPARC VIIIfx 2.0GHz Tofu Interconnect Fujitsu	705024	10510 (11280)	12659.9

TaihuLightの登場

Top 500 2017 Nov.

Name	Development	Hardware	Cores	Performance TFLOPS	Power（KW)
TaihuLight(太湖之光)	National Supercomputing Center in Wuxi	ShinWei(神威) NRCPC	10649600	93014.6 (125435.9)	15371
Tianhe-2(天河) (China)	National University of Defence Technology	Intel Xeon E5-2692 12C 2.2GHz,TH Express-2, Intel Xeon Phi31S1P	3120000	33862.7 (54902.4)	17808
Piz Daint (Switzerland)	Swiss National Supercomputing Centre	Cray XC50, Xeon E5-2690v3 12C 3.6GHz, Aries Internconnect, NVIDIA Tesla P100	361760	19590 (25326)	2272
Gyoukou (Japan)	JAMEST	ZettaScaler 2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700MHz	19860000	19135.8 (28192)	1350
Titan (USA)	DOE/SC/Oak Ridge National Lab.	Cray XK7, Opteron 6274 16C 2.2GHz, Cray Gemini Intercon.NVIDIA K20x	550640	17590 (27112.5)	8209

3位以降の変動

Green 500 2014 Nov.

	Machine	Place	FLOPS/W	Total kW
1	L-CSC, Intel Xeon E5+AMD FirePro	GSI Helmholtz Center	5271.81	57.15
2	Suiren, Xeron E5+PEZY-SC	KEK（高エネ研）	4945.63	37.83
3	TSUBAME-KFC, Intel Xeon E5+ NVIDIA K20x, Infiniband FDR	Tokyo Institute of Technology	4447.58	35.39
4	Storm 1 Xeon E5+ NVIDIA K20	Cray Inc.	3962.73	44.54
5	Wilkes Dell T620 Cluster, Intel Xeon E5+NVIDIA K20, Infiniband FDR	Cambridge University	3631.70	52.62

多くはNVIDIA Kepler K20 GPUsを利用,
PEZY-SC は独自プロセッサ

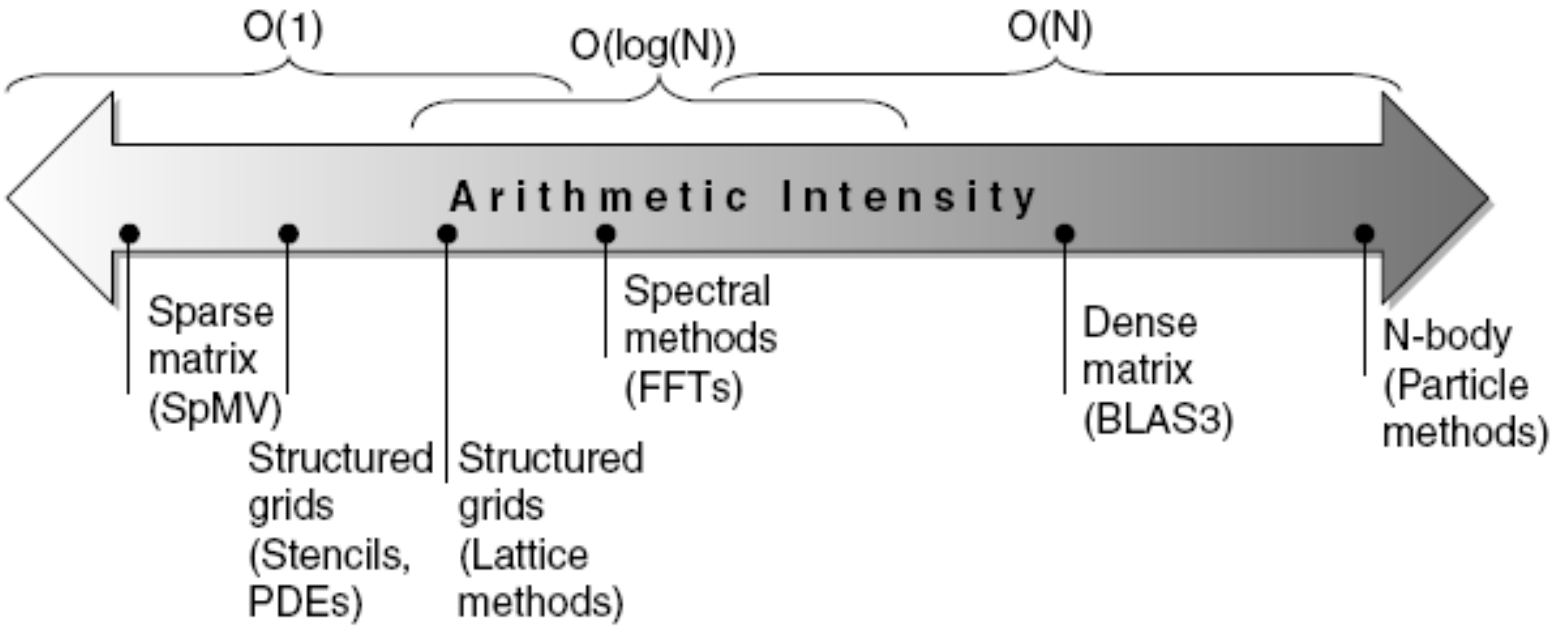
Green 500 2017 Nov.

rank(top500)	Machine	Place	GFLOPS/W	Total kW
1 （259）	Shoubu system B PEZY-SC2	RIKEN	17.009	50
2 （307）	Suiren2 PEZY-SC2	KEK（高エネ研）	16.759	47
3 （276）	Sakura PEZY-SC2	Pezy Computing	16.657	50
4 （149）	DGX SaturnV Volta NVIDIA Tesla V100	NVIDIA Corporation	15.117	97
5 （4）	Gyoukou PEZY-SC2	JAMESAT	14.173	1350

1－3、5位はPEZY-SC2利用、4位はNVIDIAのVolta

計算強度

- 読み出したデータ量（byte)に対する浮動小数点演算数

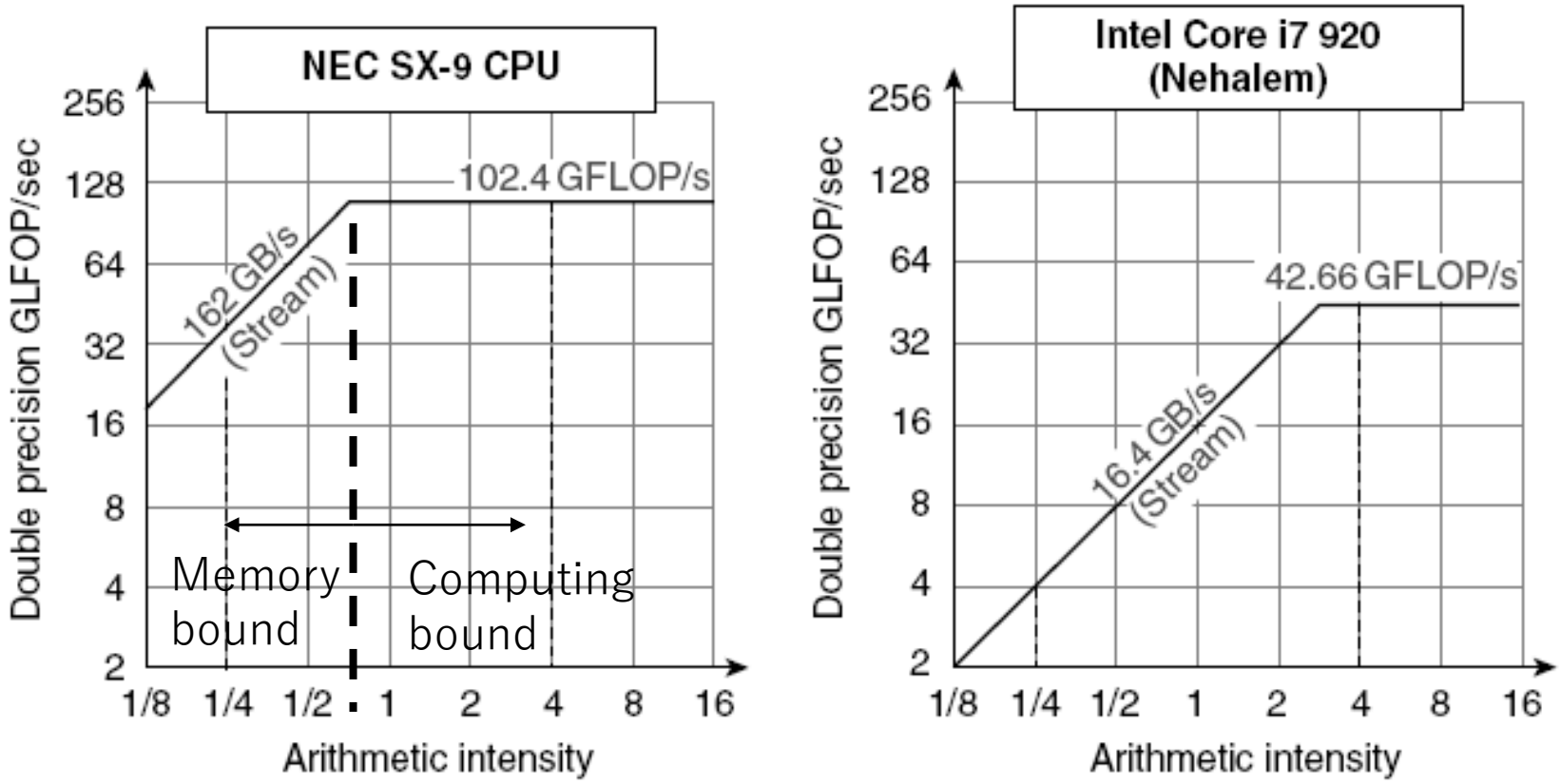


From Hennessy & Patterson’s Texbook

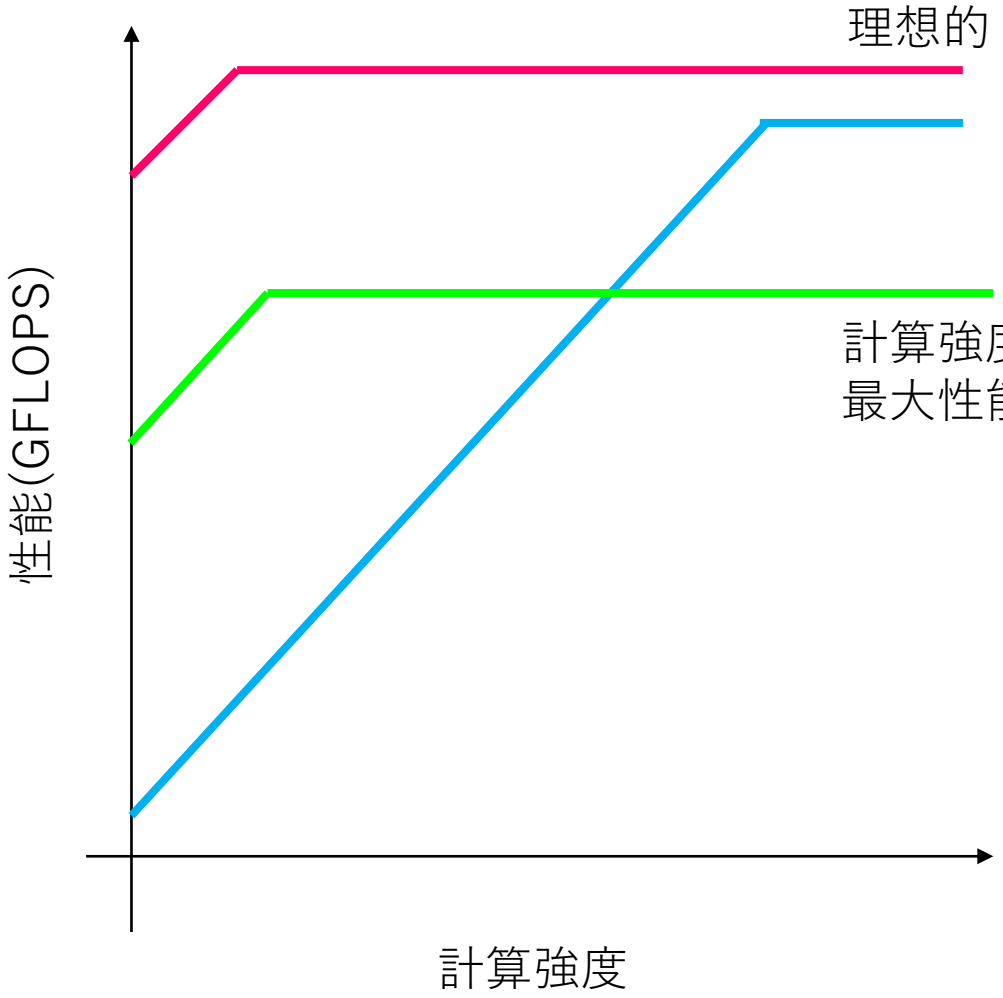
ルーフラインモデル

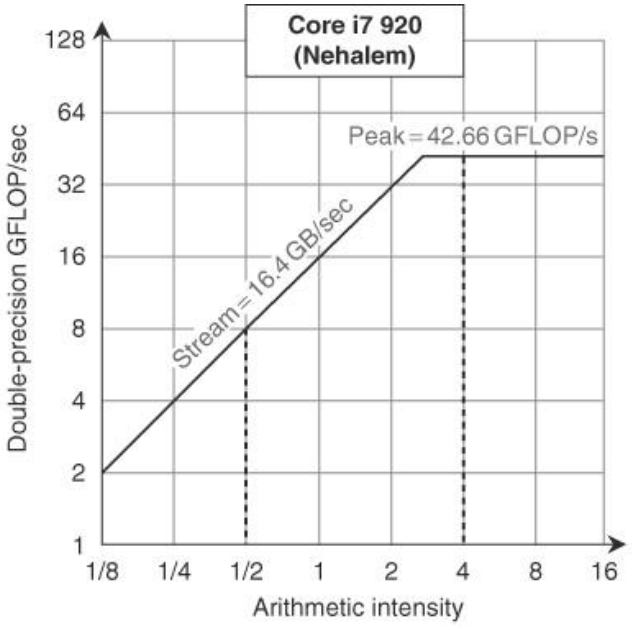
- 計算強度に対して性能を見る

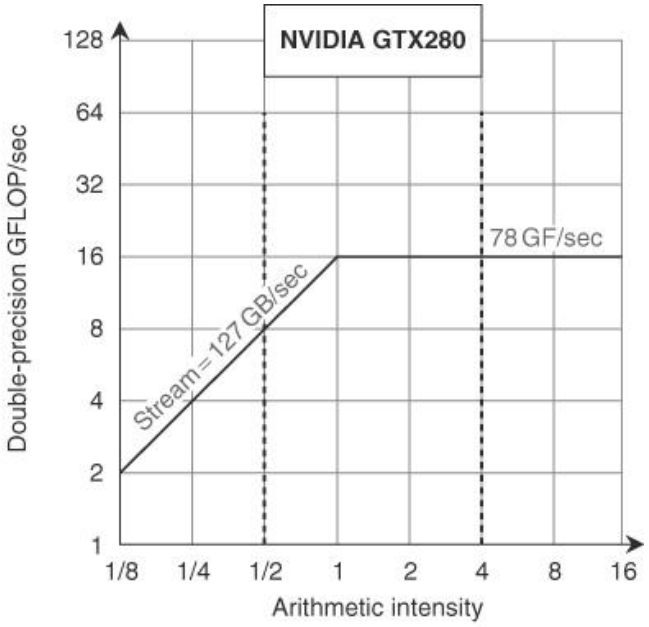
From Hennessy & Patterson's Texbook



どちらが良いコンピュータか？

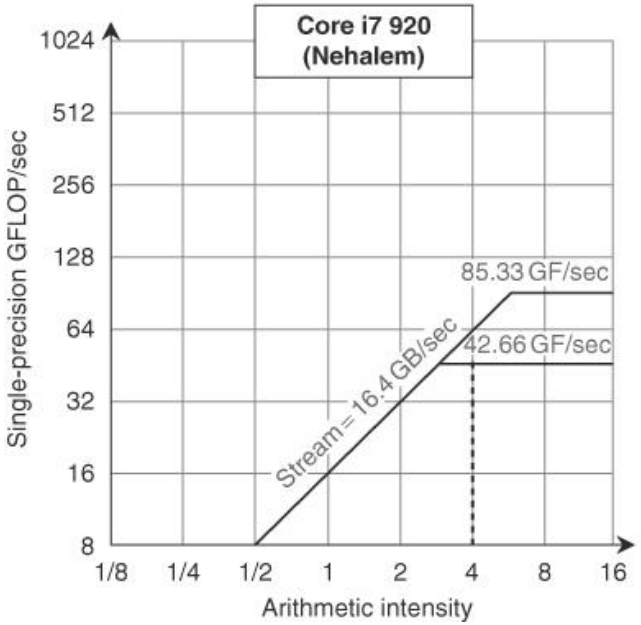


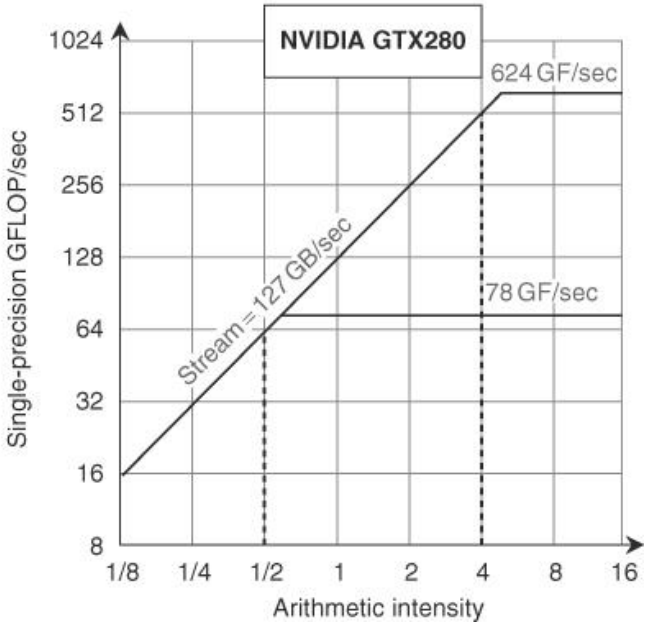




色々なコンピュータ
のルーフラインモデル

from Hennessy&
Patterson's Textbox





1位じゃなければだめなんですか？（2011年）

- 民主党政権が誕生し、無駄な公共事業を見直す「事業仕分け」の一つにスーパーコンピュータ「K：京」が取り上げられた。
- Top500の1位になることにどれだけ意味があるのか？
  - 紹介済だが、Linpackの計算性能が1位になっても技術的な価値がさほど高いわけではない。
  - しかし、過去において1位を取ったスーパーコンピュータは他の性能も悪くなく、一定の期間1位に君臨した。
  - 1位になると宣伝効果が違う！→これはむしろマスコミに問題がある。
  - 世界のトップレベルのスーパーコンピュータを持つことは、周辺の高い技術を保持することである。
  - 日本の純粋な平和利用科学目的のスーパーコンピュータが人類の科学技術、医学、生物学、天文学….に貢献する。
  - スーパーコンピュータはテクノロジにおける戦略物資でありこれを放棄することはできない。

  などの意見でこの事業仕分けは止めになった。
- 「K」は加速予算を得て予定通り1位を奪取し、長年にわたってTop5以内を維持した
  - 稼働率が高く、様々な公募プロジェクトに利用されている
- スーパーコンピュータが政治問題化した最初の事例

スパコン研究費詐欺事件に対する技術的コメント（2017年）

- 低消費電力スパコンを開発してGreen500を独占し、Top500の4位に食い込んだベンチャービジネスPezyの社長が研究費をめぐる詐欺事件で逮捕された。週刊誌、新聞はPezyの欠陥スパコンに多すぎる開発費用が国費から投じられたと報じた。→詐欺事件の本質についてはわからない。ここでは純粋な技術論を紹介する。
- Pezyのコンピュータは「欠陥スパコン」
  - 熱暴走でハングアップする、と書かれていたがこれは明らかにおかしい
    - 液浸技術を用いており、システム全体を液体に沈めてしまう
    - 熱暴走する危険性は極めて小さい
    - もちろん開発途上のコンピュータは様々な理由でハングアップするので、Gyoukouも当初は頻繁にしたとは思うがそれを持って欠陥とはいえない
  - Top500でLinpackの結果が採用されたということは一応の信頼性はあるはず
  - 以下の疑問はある
    - 独自の命令セットを用いているため、ソフトウェアの開発は困難
      - これを言い出すとARM、RISC-Vの軍門に降る以外になくなる
    - 長期間きちんとした運用ができるのか？
  - そもそもスパコンは一定の期間を運用してからその業績を評価すべき。Top500で1位とか騒ぐ方が間違っている→これはマスコミに問題がある
- Pezyのコンピュータに100億円も高額な予算をつぎ込んだ
  - 世界のトップ４、日本のトップ１を取るには通常一桁多めの予算が必要。
  - 100億円は安すぎる→どこかに問題があるのでは？との検証は必要だったかも
  - そもそもスーパーコンピュータは国家予算をつぎ込まないと開発できない。
    - どのような種類のスパコンにどの程度つぎ込むか？は議論が必要
    - この議論がきちんとなされていたか、ということには疑問がある
    - 日本のコンピュータ系の予算はスパコンにつぎ込まれ過ぎな部分もあるがこれは別の問題

演習　今日はエッセイ

- 日本は京の100倍の性能を持つエクサスケールコンピュータを2021年稼働を目標に開発している。
- 以下のそれぞれのステップに対してどのように考えるかを、それぞれ数行で簡単に書きなさい。

Step1:　エクサスケールコンピュータを税金を投じて開発する必要があるのかどうか、考えを述べなさい。

Step2:　あなたが開発するとしたら、ホモジーニアス型にしますか？アクセラレータ型にしますか？またどの理由はどのようなものですか？

Step3:　エクサスケールコンピュータが2011年に世界一になったKにつづいて2021年にTop500の世界一を奪取したとしましょう。2031年にこの100倍の性能を持つスーパーコンピュータを開発すべきだと思いますか？またその理由はなぜですか？