

コンピュータ構成と設計 (13)

性能評価と高性能コンピューティング

李 亜民

2025 年 1 月 9 日 (木)

コンピュータの性能

- コンピュータの性能

$$= \frac{1}{\text{プログラムを実行する際にかかる時間}}$$

- 応答時間 (レイテンシまたはレスポンスタイム)

- ▶ 一つのジョブ¹を実行する際にかかる時間

- スループット

- ▶ 単位時間内にシステムによって処理されるジョブ量

¹ 計算機に処理させるひとまとまりの仕事の単位

実行時間

- 経過時間

- ▶ すべての時間を計る (ディスクアクセス, メモリアクセス, I/O)
- ▶ 役に立つ値だが, 比較目的には使いにくい場合が多い

- CPU 時間

- ▶ I/O や他のプログラムの実行に掛かる時間は計らない
- ▶ システム CPU 時間とユーザ CPU 時間に分けることができる

- ユーザ CPU 時間

- ▶ 自分のプログラムに書かれているコードを実行するのに掛かる時間

実行時間

- 性能 = $\frac{1}{\text{実行時間}}$
- 実行時間 = $I \times \text{CPI} \times \text{TPC}$
 - ▶ I: 実行命令数
 - ▶ CPI: 命令あたりの平均クロック・サイクル数
 - ▶ TPC: クロック・サイクル時間
- クロック周波数 F
 - ▶ $\text{TPC} = \frac{1}{F}$

例題

● 問題

- ▶ 計算機 A: $\text{CPI}_A = 4$, $F_A = 2\text{GHz}$
- ▶ 計算機 B: $\text{CPI}_B = 3$, $F_B = 1.8\text{GHz}$
- ▶ どちらが優れているか？

例題

● 問題

- ▶ 計算機 A: $CPI_A = 4, F_A = 2\text{GHz}$
- ▶ 計算機 B: $CPI_B = 3, F_B = 1.8\text{GHz}$
- ▶ どちらが優れているか？

● 解答

- ▶ 実行時間_A = $I \times 4 \times (1/2)$
- ▶ 実行時間_B = $I \times 3 \times (1/1.8)$
- ▶ $\frac{\text{性能}_A}{\text{性能}_B} = \frac{\text{実行時間}_B}{\text{実行時間}_A} = \frac{I \times 3 \times (1/1.8)}{I \times 4 \times (1/2)} = \frac{3 \times 2}{4 \times 1.8} = \frac{1}{1.2} < 1$
- ▶ 計算機 B の方が優れている

MIPS: Million Instructions Per Second

- MIPS = 1 秒間に処理できる百万命令数 =

$$\frac{\text{実行命令数}}{\text{実行時間} \times 10^6}$$

- 問題

- ▶ 計算機 A: $\text{CPI}_A = 4$, $F_A = 2\text{GHz}$, $\text{MIPS}_A = ?$
- ▶ 計算機 B: $\text{CPI}_B = 3$, $F_B = 1.8\text{GHz}$, $\text{MIPS}_B = ?$

MIPS: Million Instructions Per Second

- MIPS = 1 秒間に処理できる百万命令数 =
$$\frac{\text{実行命令数}}{\text{実行時間} \times 10^6}$$

- 問題

- ▶ 計算機 A: $\text{CPI}_A = 4$, $F_A = 2\text{GHz}$, $\text{MIPS}_A = ?$
- ▶ 計算機 B: $\text{CPI}_B = 3$, $F_B = 1.8\text{GHz}$, $\text{MIPS}_B = ?$

- 解答

- ▶
$$\text{MIPS}_A = \frac{I}{I \times \text{CPI} \times \frac{1}{F_A} \times 10^6} = \frac{F_A}{\text{CPI} \times 10^6} = \frac{2 \times 10^9}{4 \times 10^6} = 500$$
- ▶
$$\text{MIPS}_B = \frac{I}{I \times \text{CPI} \times \frac{1}{F_B} \times 10^6} = \frac{F_B}{\text{CPI} \times 10^6} = \frac{1.8 \times 10^9}{3 \times 10^6} = 600$$

● 問題

サイクル数:	1	2	3	4
プログラム A:	50%	25%	10%	15%
プログラム B:	40%	30%	20%	10%

▶ $CPI_A = ?$

▶ $CPI_B = ?$

● 問題

サイクル数:	1	2	3	4
プログラム A:	50%	25%	10%	15%
プログラム B:	40%	30%	20%	10%

▶ $CPI_A = ?$

▶ $CPI_B = ?$

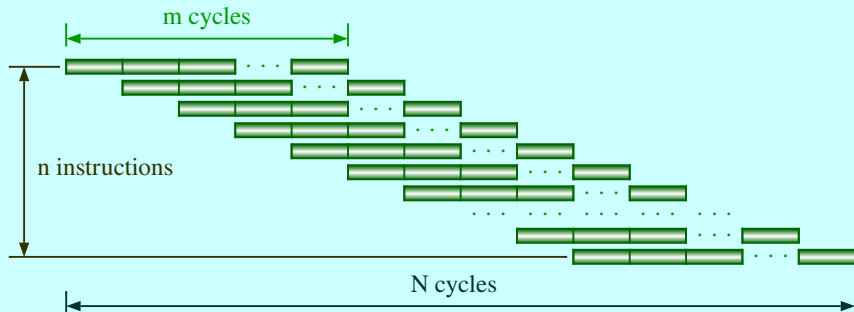
● 解答:

▶ $CPI_A = 1 \times 0.5 + 2 \times 0.25 + 3 \times 0.1 + 4 \times 0.15 = 1.9$

▶ $CPI_B = 1 \times 0.4 + 2 \times 0.3 + 3 \times 0.2 + 4 \times 0.1 = 2.0$

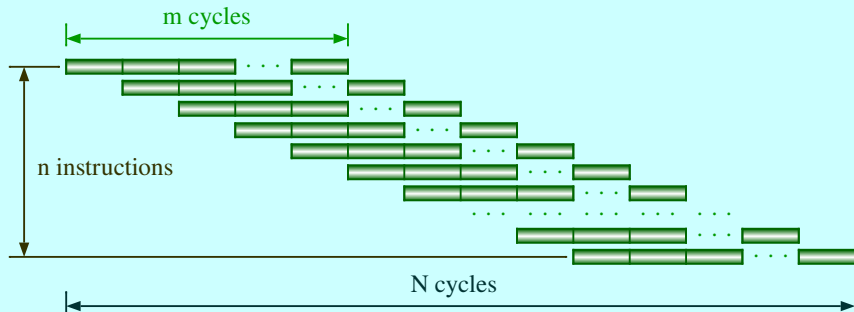
パイプラインCPU

- 問題: $N = f(m, n) = ?$



パイプラインCPU

- 問題: $N = f(m, n) = ?$



- 解答:

$$N = n + (m - 1) = n + m - 1$$

$$N = m + (n - 1) = n + m - 1$$

高性能コンピュータ

- 高性能 CPU

- ① スーパースカラ (Superscalar) CPU
- ② マルチスレッディング (Multithreading) CPU
- ③ マルチコア (Multi-Core) CPU

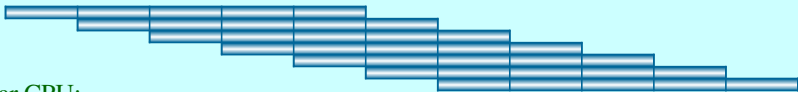
- 並列コンピューティング (Parallel Computing)
複数のマイクロプロセッサやコンピュータに処理を分散して割り当て、同時に計算・処理を行なうことで、システム全体の処理性能を向上させる技術の総称

- ① スーパーコンピュータ (Supercomputer)
- ② 相互結合網 (Interconnection Networks)

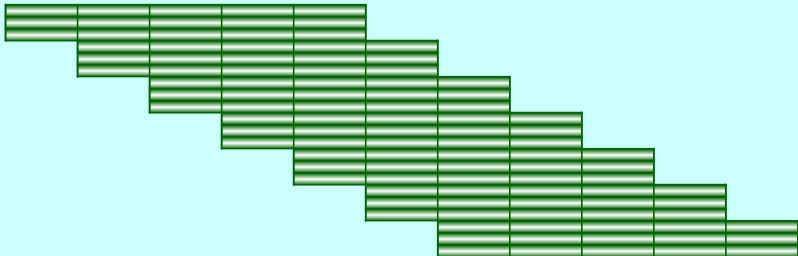
スーパースカラ CPU

スーパースカラ (Superscalar) — プロセッサの中に複数の処理系統 (パイプライン) を用意し、複数の命令を並列に処理すること

Pipelined CPU:

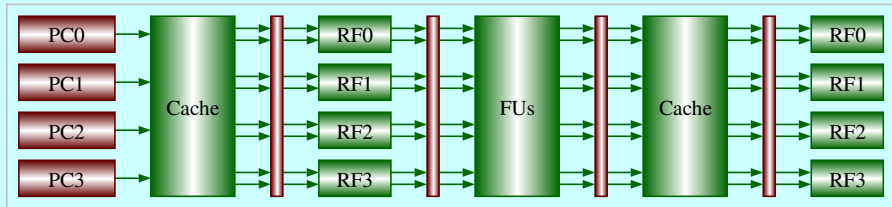


Superscalar CPU:



マルチスレッディング CPU

マルチスレッディング (Multithreading) とは、単一 CPU により複数のスレッド (コンピュータプログラミング) を同時に実行するプロセッサの機能



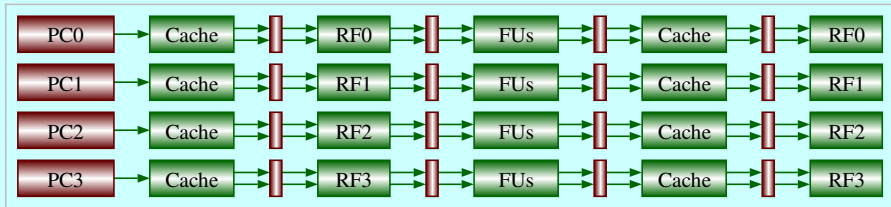
PC — Program Counter

FUs — Functional Units (ALUs, FPUs, ...)

RF — Register File

マルチコア CPU

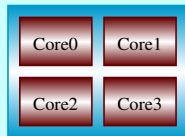
マルチコア (Multi-Core) は、1つのプロセッサ・パッケージ内に複数のプロセッサ・コアを封入した技術であり、マルチ・プロセッシングの一形態である



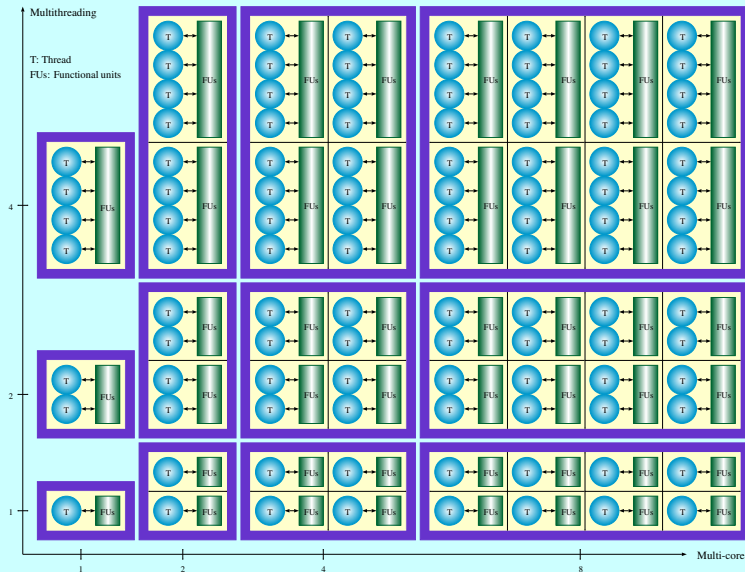
PC — Program Counter

FUs — Functional Units (ALUs, FPUs, ...)

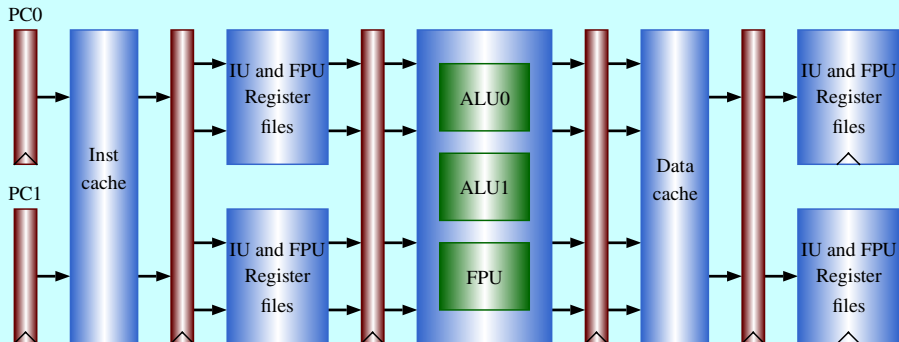
RF — Register File



マルチコアマルチスレッディングCPU

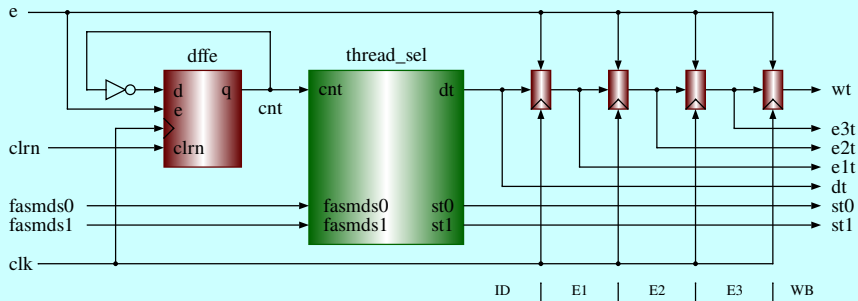


Block Diagram of 2-Thread CPU



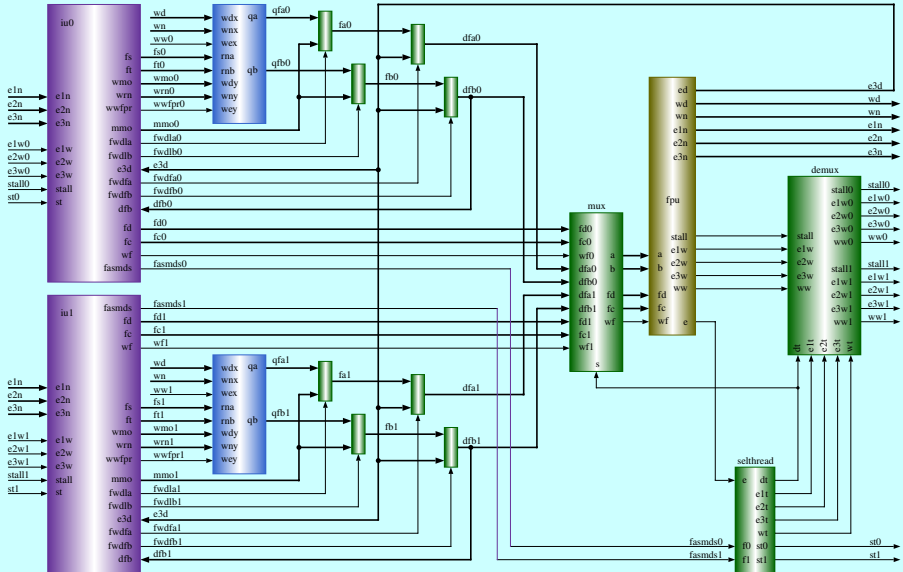
Two threads share functional units

Thread Selection Method

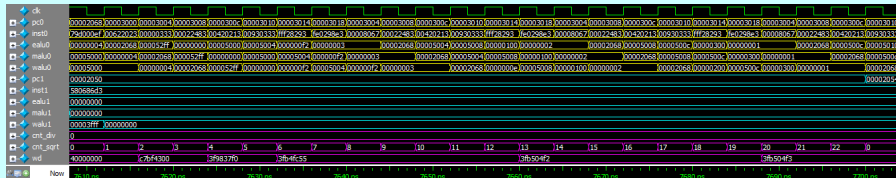


cnt	fasmds0	fasmds1	dt	st0	st1
0	1	0	0	0	0
0	1	1	0	0	1
0	0	1	1	0	0
1	0	1	1	0	0
1	1	1	1	1	0
1	1	0	0	0	0
x	0	0	0	0	0

Multithreading CPU with 2-IU 1-FPU



Waveform of Multithreading CPU

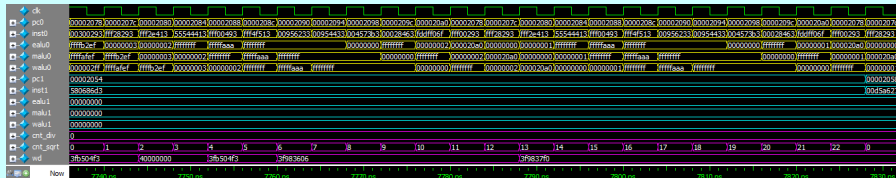


pc1: # pc1: PC of thread 1

```

2040: flw      fa0, 4(a1)    # (40800000)
2044: flw      fa1, 8(a1)    # (40000000)
2048: fdiv.s   fa3, fa0, fa1 # (40800000) / (40000000) = (40000000)
204c: fsqrt.s  fa3, fa3      # (40000000)      sqrt    = (3fb504f3)
2050: fsqrt.s  fa3, fa3      # (3fb504f3)      sqrt    = (3f9837f0)
2054: fsw       fa3, 12(a1)   # store 3f9837f0 to memory
    
```

Waveform of Multithreading CPU

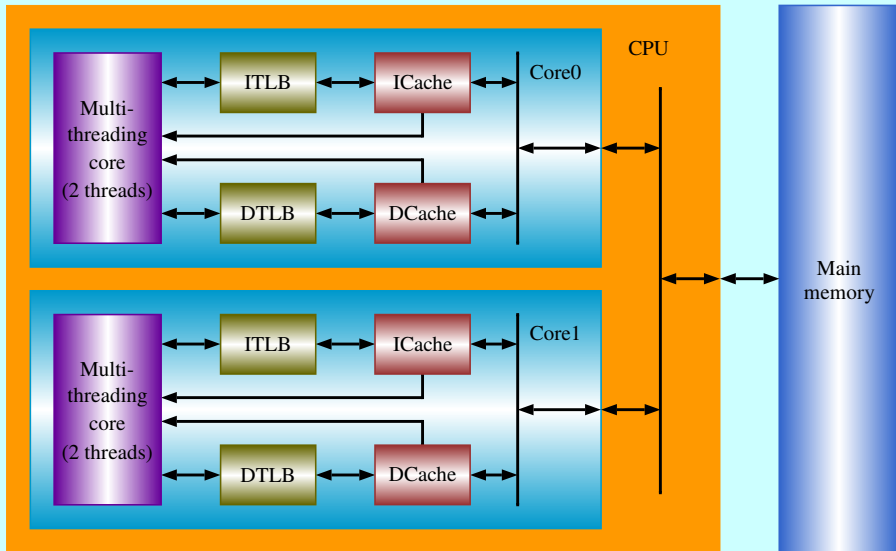


pc1: # pc1: PC of thread 1

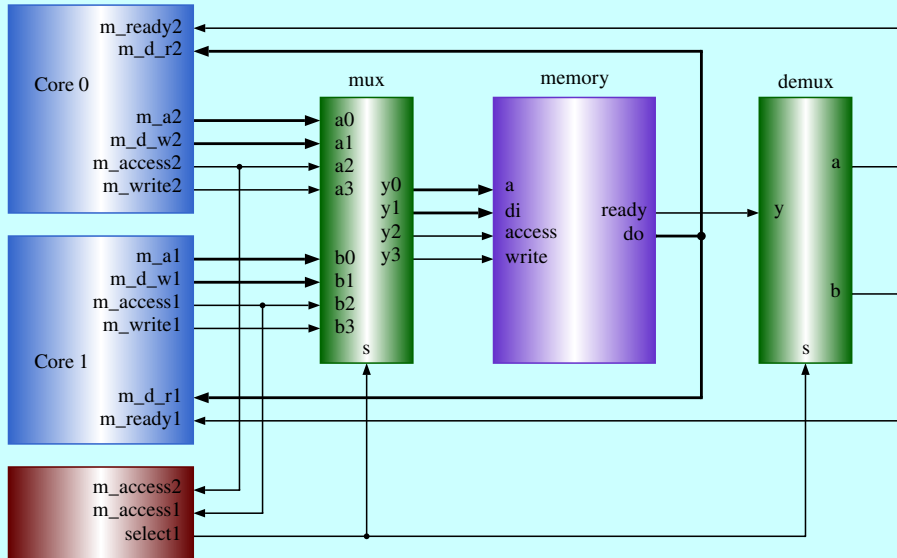
```

2040: flw      fa0, 4(a1)    # (40800000)
2044: flw      fa1, 8(a1)    # (40000000)
2048: fdiv.s   fa3, fa0, fa1 # (40800000) / (40000000) = (40000000)
204c: fsqrt.s  fa3, fa3       # (40000000)      sqrt    = (3fb504f3)
2050: fsqrt.s  fa3, fa3       # (3fb504f3)      sqrt    = (3f9837f0)
2054: fsw      fa3, 12(a1)    # store 3f9837f0 to memory
    
```

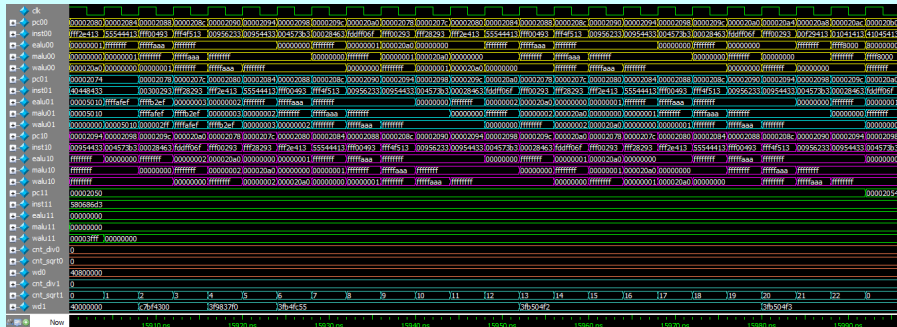
Multicore Multithreading CPU Design



Memory Access Arbitration



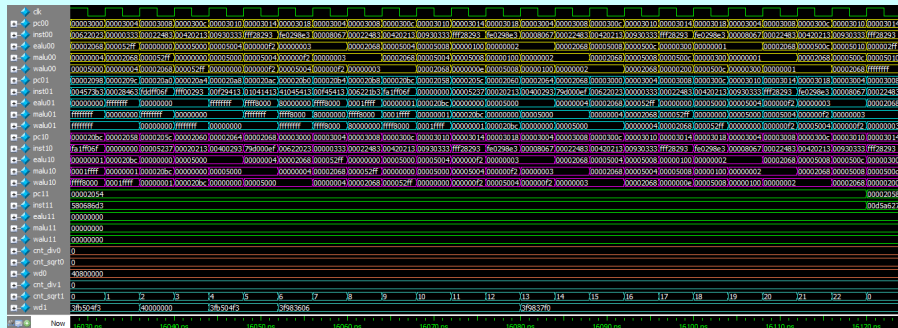
Waveform of 2-Core 4-thread CPU



```

pc11:                                     # pc11: PC of core 1 thread 1
2040: flw      fa0, 4(a1)                  # (40800000)
2044: flw      fa1, 8(a1)                  # (40000000)
2048: fdiv.s   fa3, fa0, fa1               # (40800000) / (40000000) = (40000000)
204c: fsqrt.s  fa3, fa3                    # (40000000)      sqrt      = (3fb504f3)
2050: fsqrt.s  fa3, fa3                    # (3fb504f3)      sqrt      = (3f9837f0)
2054: fsw      fa3, 12(a1)                 # store 3f9837f0 to memory
    
```

Waveform of 2-Core 4-thread CPU

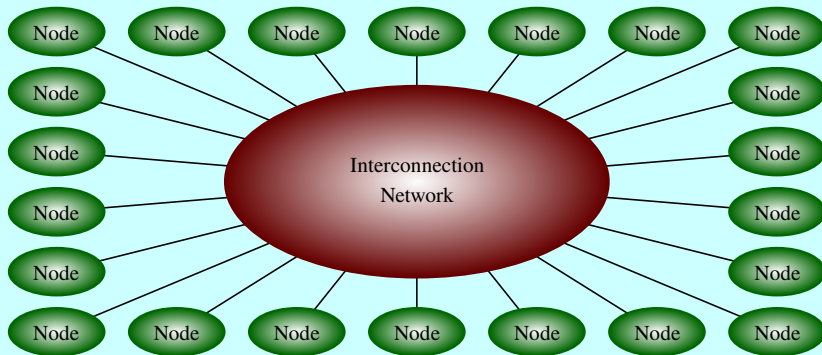


```

pc11:                                     # pc11: PC of core 1 thread 1
2040: flw      fa0, 4(a1)                  # (40800000)
2044: flw      fa1, 8(a1)                  # (40000000)
2048: fdiv.s   fa3, fa0, fa1               # (40800000) / (40000000) = (40000000)
204c: fsqrt.s  fa3, fa3                    # (40000000)      sqrt      = (3fb504f3)
2050: fsqrt.s  fa3, fa3                    # (3fb504f3)      sqrt      = (3f9837f0)
2054: fsw      fa3, 12(a1)                 # store 3f9837f0 to memory
    
```

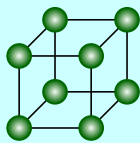
スーパーコンピュータ

スーパーコンピュータ（Supercomputer）とは、内部の演算処理速度がその時代の一般的なコンピュータより非常に高速な計算機（コンピュータ）のこと

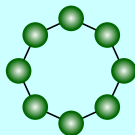


相互接続網 (ネットワーク)

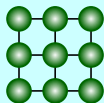
相互結合網 (Interconnection Networks) は並列計算機を構成する複数のプロセッサやメモリを相互に結合し、その間における通信路を提供する



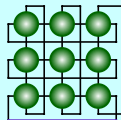
Hypercube



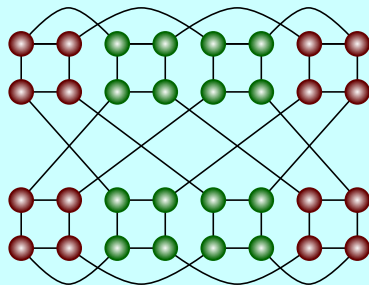
Ring



2D Mesh

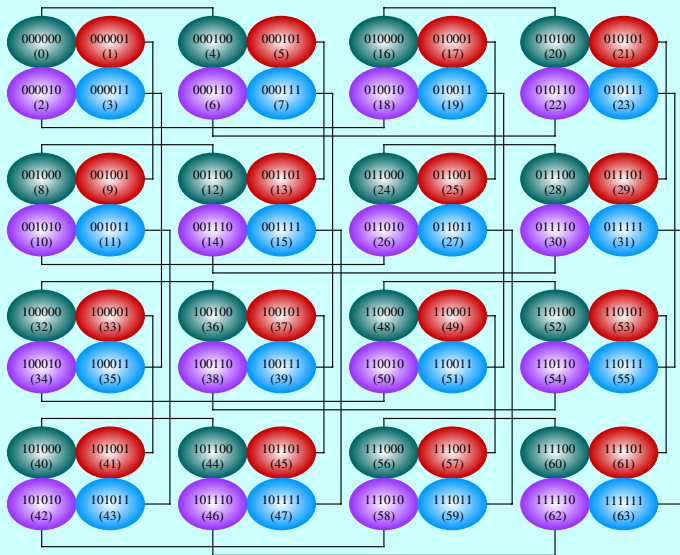


2D Torus

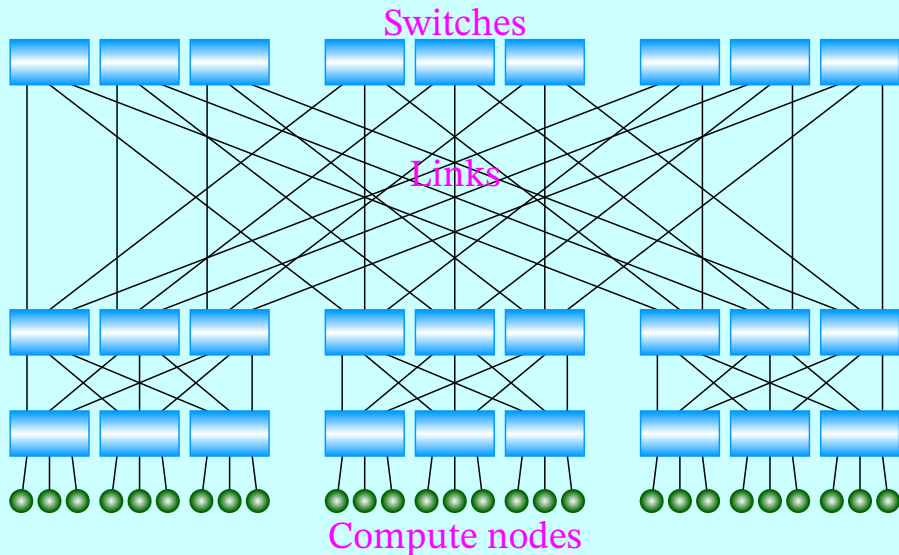


Dual-Cube

相互接続網 (Metacube)



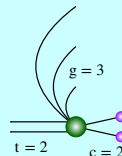
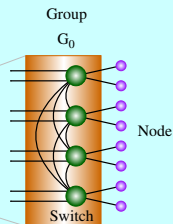
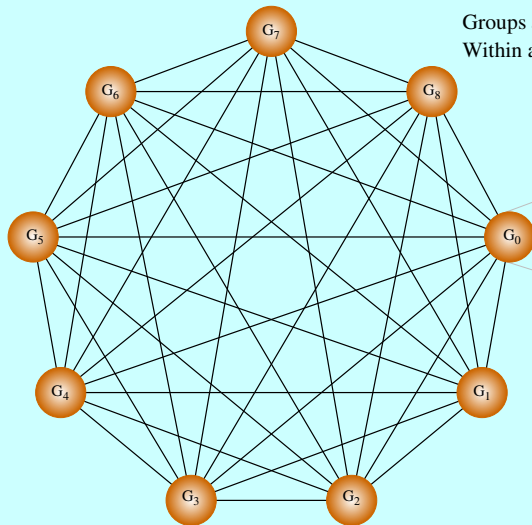
相互接続網 (Fat-Tree)



相互接続網 (Dragonfly)

Groups are fully connected.

Within a group: switches are fully connected.

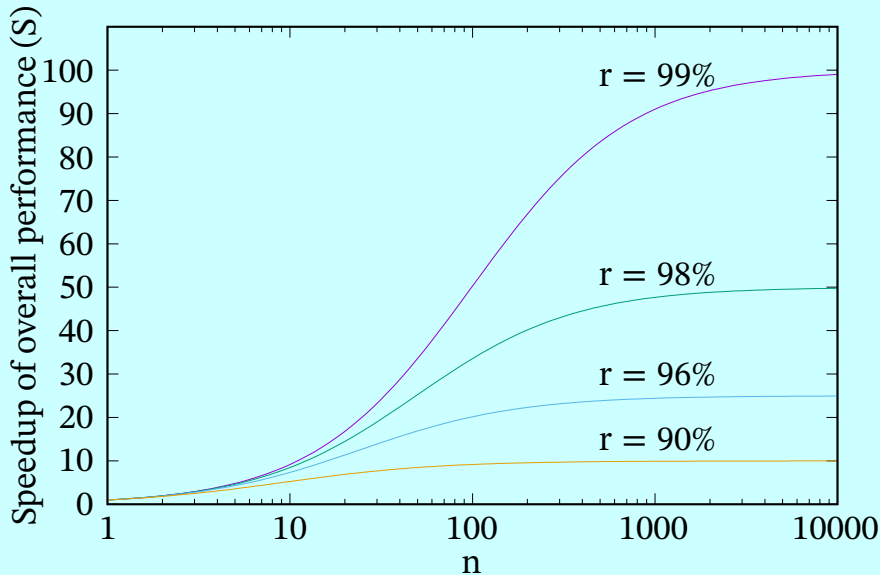


アムダールの法則 (Amdahl's Law)

- Amdahl's Law states that the performance improvement to be gained from using some faster mode of execution is limited by the fraction of the time the faster mode can be used.
- Let P_n be the performance with enhancement, P_o be the performance without enhancement, T_n be the execution time with enhancement, T_o be the execution time without enhancement, r be the fraction enhanced, and n be the speedup of the enhanced section. Speedup

$$S = \frac{P_n}{P_o} = \frac{T_o}{T_n} = \frac{T_o}{T_o \times r/n + T_o \times (1 - r)} = \frac{1}{r/n + (1 - r)}$$

アムダールの法則 (Amdahl's Law)



Top 500 Supercomputers

TOP500 Supercomputer Sites: <https://www.top500.org/>

TOP 10 Supercomputers in November 2024

Rank	System	Country	Maker	Cores	PFlop/s
1	El Capitan	USA	HPE	11,039,616	1,742.00
2	Frontier	USA	HPE	9,066,176	1,353.00
3	Aurora	USA	Intel	9,264,128	1,012.00
4	Eagle	USA	Microsoft	2,073,600	561.20
5	HPC6	Italy	HPE	3,143,520	477.90
6	Fugaku	Japan	Fujitsu	7,630,848	442.01
7	Alps	Switzerland	HPE	2,121,600	434.90
8	LUMI	Finland	HPE	2,752,704	379.70
9	Leonardo	Italy	EVIDEN	1,824,768	241.20
10	Tuolumne	USA	HPE	1,161,216	208.10

課題 XIII (100 点 + 100 点)

- 1 Explain the Superscalar, Multithreading, Multicore, Supercomputer, and Interconnection Networks.
- 2 Let $n = 10$ and $r = 75\%$. Calculate the overall speedup S by using Amdahl's Law and give the upper bound of S .
- 3 Let $n = 1,000,000$ and $S = 500,000$. Use Amdahl's law to calculate fractions r and $1 - r$. What do these numbers mean when comparing supercomputer performance to uniprocessor computer performance?

課題 XIII (100 点 + 100 点)

- ④ Option (+100 点) : Design an RISC-V multicore multithreading (2×2) CPU RV32IMF that contains instruction cache & TLB and data cache & TLB.