Chapter 3

Performance Analysis of Multiprocessor Architecture

3.1 Computational Models

- Equal Duration Model
  - A task is divided into n equal subtasks each of which is executed by 1 processor.
  - t_s is the execution time of the whole task using a single processor.
  - The time taken by each processor to execute its subtask $t_m = t_s/n$.
  - Then time taken by the whole system to execute the whole task is $t_m = t_s/n$.

3.1 Computational Models

- Speedup factor $S(n) = t_s/t_m = t_s/ (t_s/n) = n$.
- Speedup factor = number of processors used, n.
- Assume time t_c is needed as overhead time for processors to communicate, therefore: $t_{m=} (t_s/n) + t_{c.}$
- Speedup (n) becomes: $n/ (1+(n * t_c/t_s))$.
- Efficiency = $speedup(n)/n = 1/ (1+(n * t_c/t_s))$.

3.1 Computational Models

- Parallel Computation With Serial Sections Model:
  - A fraction f of a given task is not dividable into concurrent subtasks.
  - (1-f) is assumed to be dividable into concurrent subtasks.
  - Therefore speedup (n) = $\dfrac{t_s}{ft_s + (1-f)\dfrac{t_s}{n}}$

    = $\dfrac{n}{1+(n-1)f}$

3.1 Computational Models

- Speedup:
  - S = Speed(new) / Speed(old)
  - S = Work/time(new) / Work/time(old)
  - S = time(old) / time(new)
  - S = time(before improvement) / time(after improvement)

3.1 Computational Models

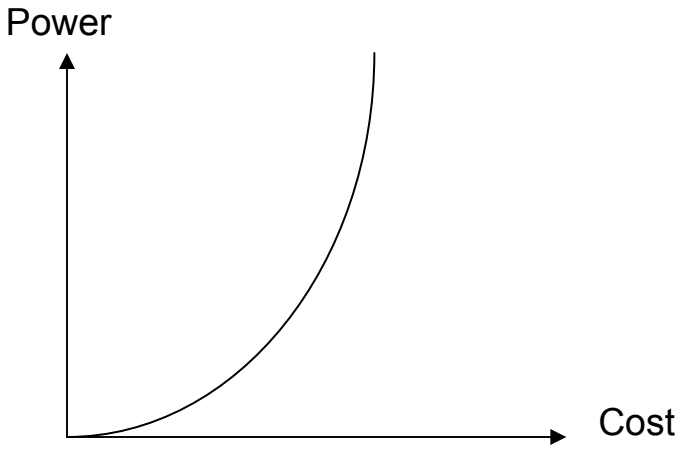
- Speedup:
  - Time (one CPU): T(1).
  - Time (n CPUs): T(n).
  - Speedup: S
  - S = T(1)/T(n)

3.2 An Argument For Parallel Architectures

- Grosch's law
  - "To sell a computer for twice as much, it must be four times as fast".
  - Vendors skip small speed improvements in favor of waiting for large ones.

3.2 An Argument For Parallel Architectures

– Buyers of expensive machines would wait for a twofold improvement in performance for the same price.

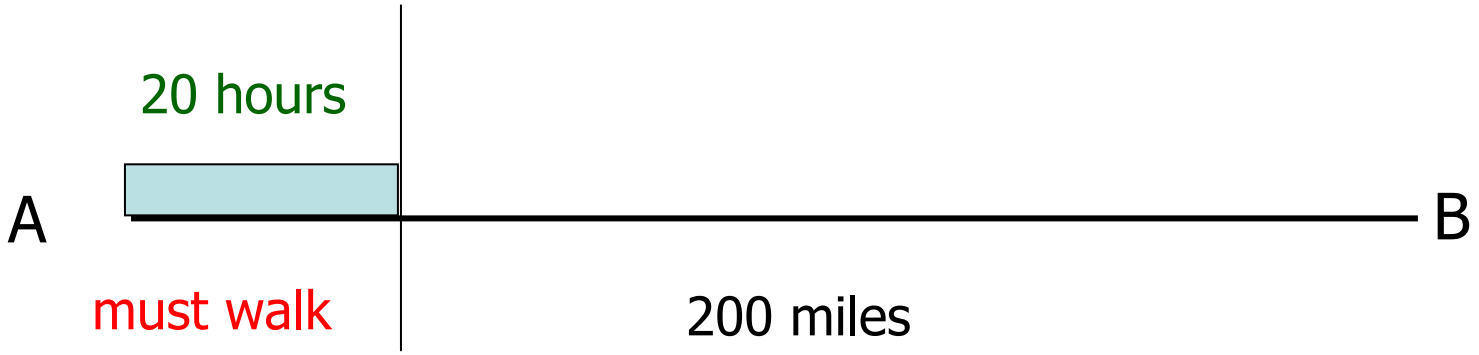


3.2 An Argument For Parallel Architectures

- Amdahl's law
  - 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.
  - There is an intrinsic limit set on the performance improvement (speed) regardless of the number of processors used.

3.2 An Argument For Parallel Architectures

- Amdahl's law



Walk 4 miles /hour	→ 50 + 20 = 70 hours	S = 1
Bike 10 miles / hour	→ 20 + 20 = 40 hours	S = 1.8
Car-1 50 miles / hour	→ 4 + 20 = 24 hours	S = 2.9
Car-2 120 miles / hour	→ 1.67 + 20 = 21.67 hours	S = 3.2
Car-3 600 miles /hour	→ 0.33 + 20 = 20.33 hours	S = 3.4

3.2 An Argument For Parallel Architectures

- Amdahl's law
  - β : The fraction of the program that is naturally serial.
  - (1- β): The fraction of the program that is naturally parallel.

3.2 An Argument For Parallel Architectures

- Amdahl's law

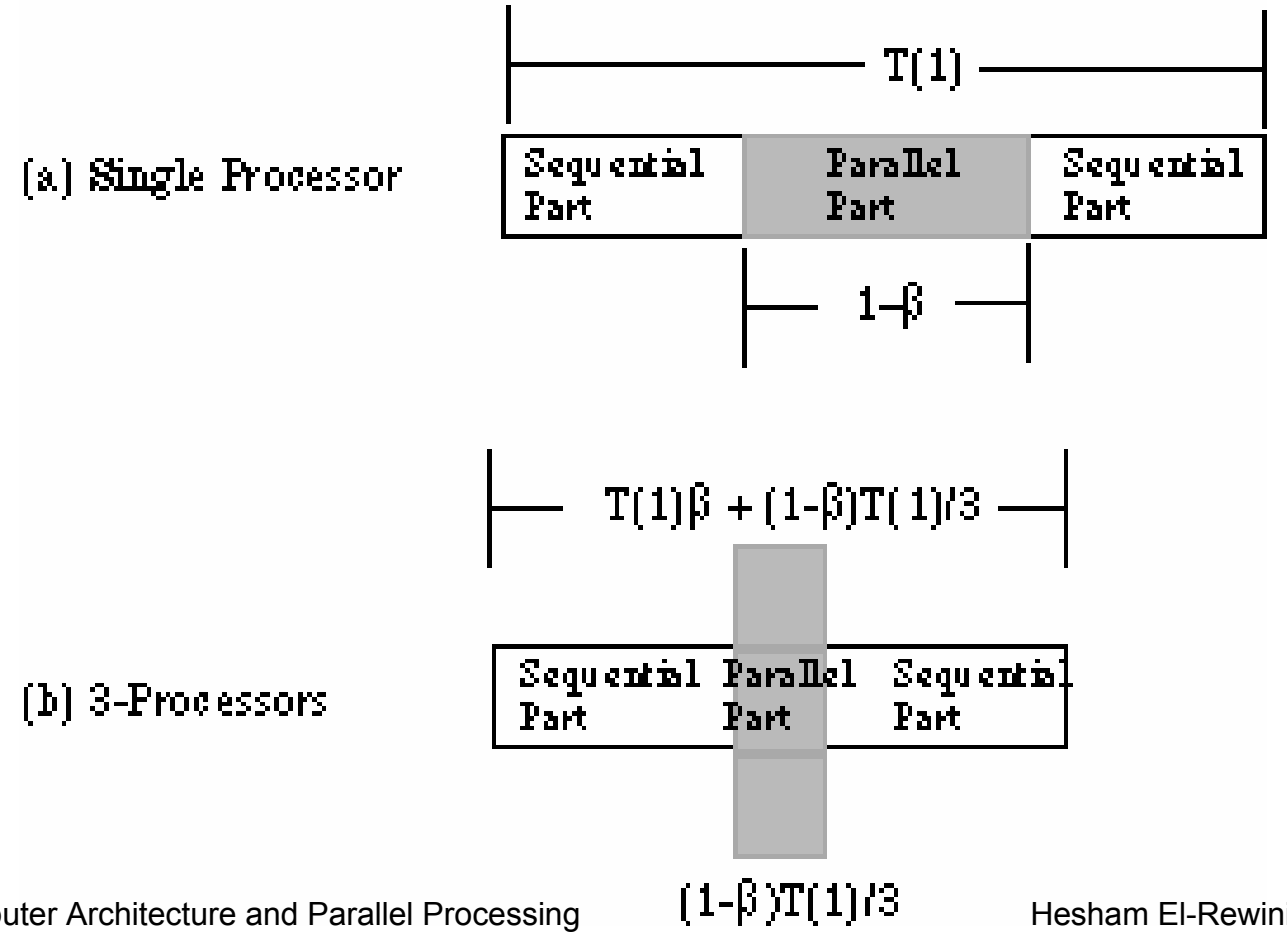
$$S = T(1)/T(N)$$

$$T(N) = T(1)\beta + \frac{T(1)(1-\beta)}{N}$$

$$S = \frac{1}{\beta + \frac{(1-\beta)}{N}} = \frac{N}{\beta N + (1-\beta)}$$

3.2 An Argument For Parallel Architectures

- Amdahl's law



3.2 An Argument For Parallel Architectures

- Gustafson-Barsis's law
  - If s and p are the serial and parallel time spent on a parallel system, then s+ p*n is the time needed by a serial processor to perform the computation.

3.2 An Argument For Parallel Architectures

- Gustafson-Barsis's law

N & β are not independent from each other.

α : The fraction of the program that is naturally serial.

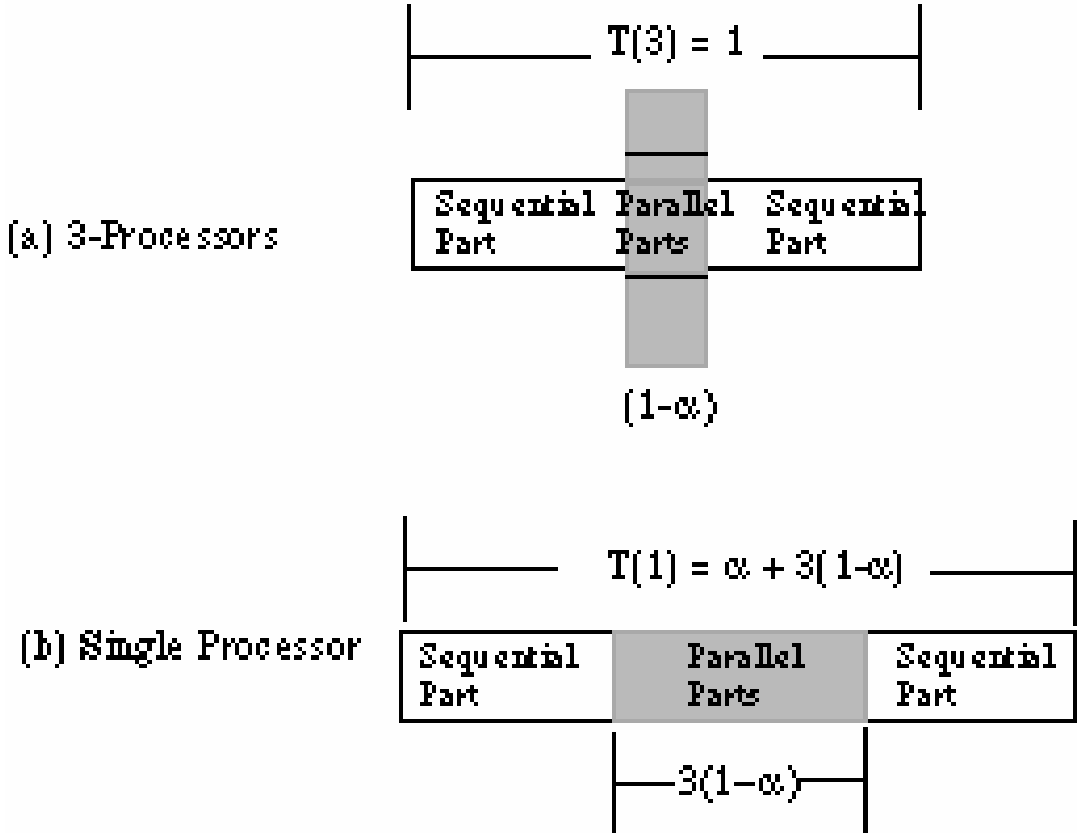
$$T(N) = 1$$

$$T(1) = \alpha + (1- \alpha) N$$

$$S = N – (N-1) \alpha$$

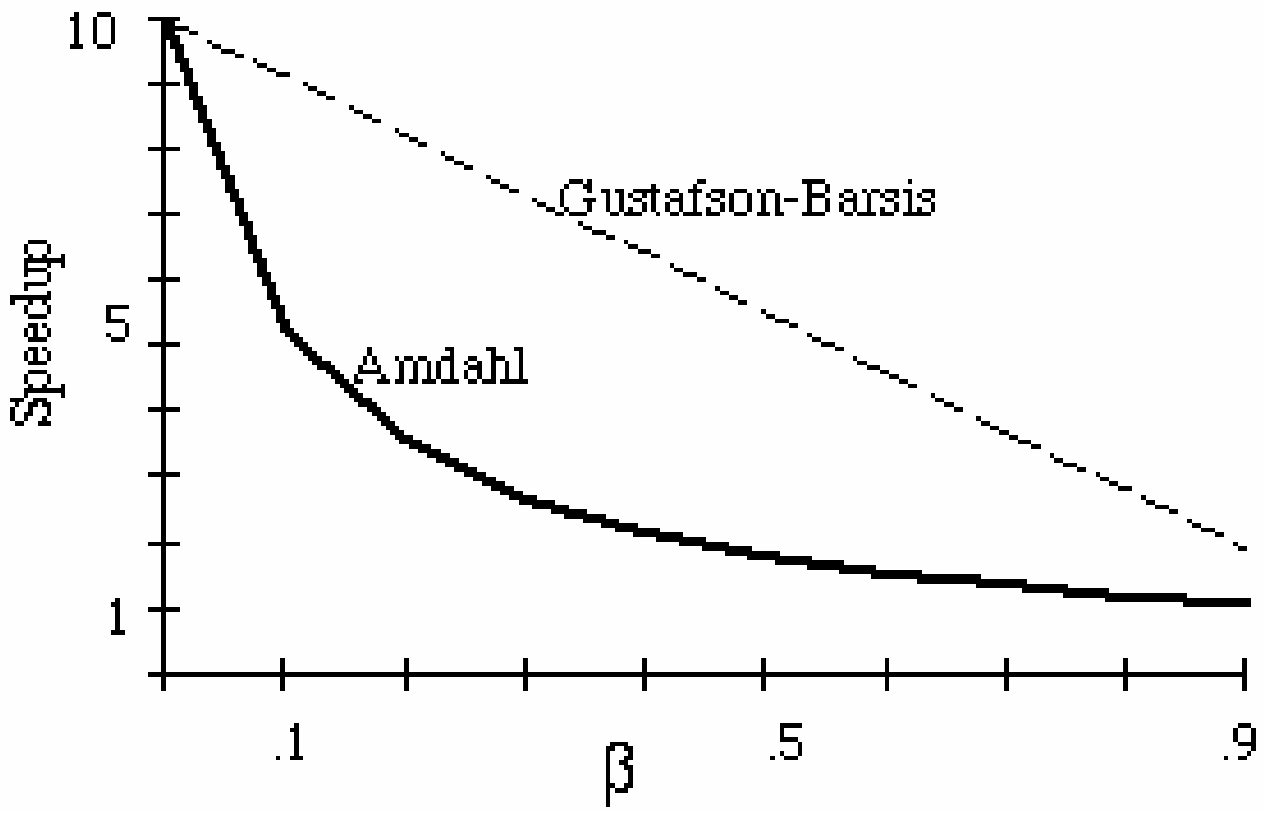
3.2 An Argument For Parallel Architectures

- Gustafson-Barsis's law



3.2 An Argument For Parallel Architectures

- Gustafson-Barsis's law



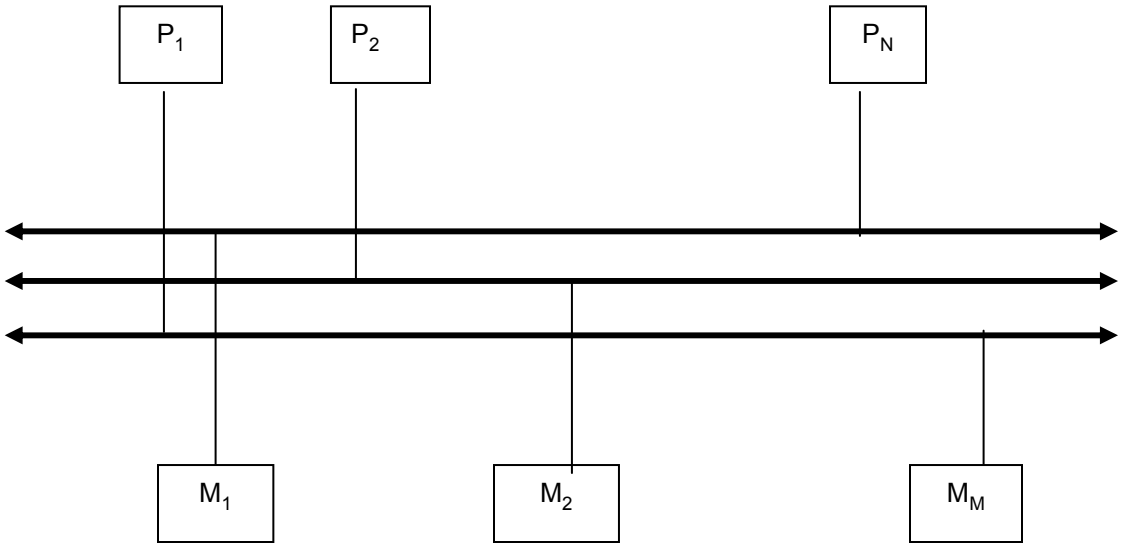
3.3 Interconnection Networks Performance Issues

- Bandwidth of a crossbar:
  - is the average number of requests that can be accepted by a crossbar in a given cycle.
  - For M memory modules and n processors,
    - if a processor generates a request with probability ρ in a cycle directed to each memory with equal probability, then the expression for the bandwidth is: $M(1-(1-(\rho/M))^n)$

3.3 Interconnection Networks Performance Issues

- Bandwidth of a multiple bus:

$$BW = \sum_{k=1}^{B} k \times \beta + \sum_{k=B+1}^{N} B \times \beta$$



3.3 Interconnection Networks Performance Issues

- Bandwidth of a Multistage Interconnection Network:
  - Assumption: MIN consists of stages of a x b crossbar switches.
  - $BW = b^n X r_n$

3.4 Scalability of Parallel Architectures

- A parallel architecture is scalable if it can be expanded (or reduced) to a larger (smaller) system with a linear increase (decrease) in its performance (cost).
- In terms of speed, a scalable system is capable of increasing its speed in proportion to the increase in number of processors.
- In terms of efficiency, a parallel system is scalable if its efficiency is kept fixed as the number of processors is increased.
- Size scalability: measures the maximum number of processors a system can accommodate.

3.4 Scalability of Parallel Architectures

- Application scalability: ability to run application software with improved performance on a scaled-up version of the system.
- Generation scalability: ability of a system to scale-up by using next-generation (fast) components.
- Heterogeneous scalability: ability of a system to scale-up by using hardware and software components supplied by different vendors.

3.5 Benchmark Performance

- Benchmark performance: is the use of a set of integer and floating-point programs (known as benchmark) designed to test different performance aspects of a computing system under test.
- Benchmark programs should be designed to provide fair and effective comparisons among high-performance computing systems.

3.5 Benchmark Performance

- Serial Benchmarks
- Parallel Benchmarks
- PERFECT Benchmarks
- NAS Kernel
- The SLALOM
- The Golden Bell Prize
- WebSTONE for the Web

3.6 Summary

- A number of issues related to the performance of multiprocessor systems was covered.
- Two computational models were introduced:
  - Equal duration
  - Parallel computations with serial sections
- A rebuttal to a number of critical views about the effectiveness of parallel architectures has been made:
  - Grosch's law
  - Amdahl's law
- A number of performance metrics for static and dynamic interconnection networks has been provided.

3.6 Summary

- The scalability of parallel architectures in terms of speed and efficiency has been discussed.
- A number of unconventional metrics for scalability has also been discussed.
- Finally, The issue of benchmark performance measurement has been introduced.