CSE-40533

Introduction to Parallel Processing

Chapter 1

- **Set the context in which the course material will be presented**
- **Review challenges that facet he designers and users of parallel computers**
- **Introduce metrics for evaluating the effectiveness of parallel systems**

1.1 Why Parallel Processing?

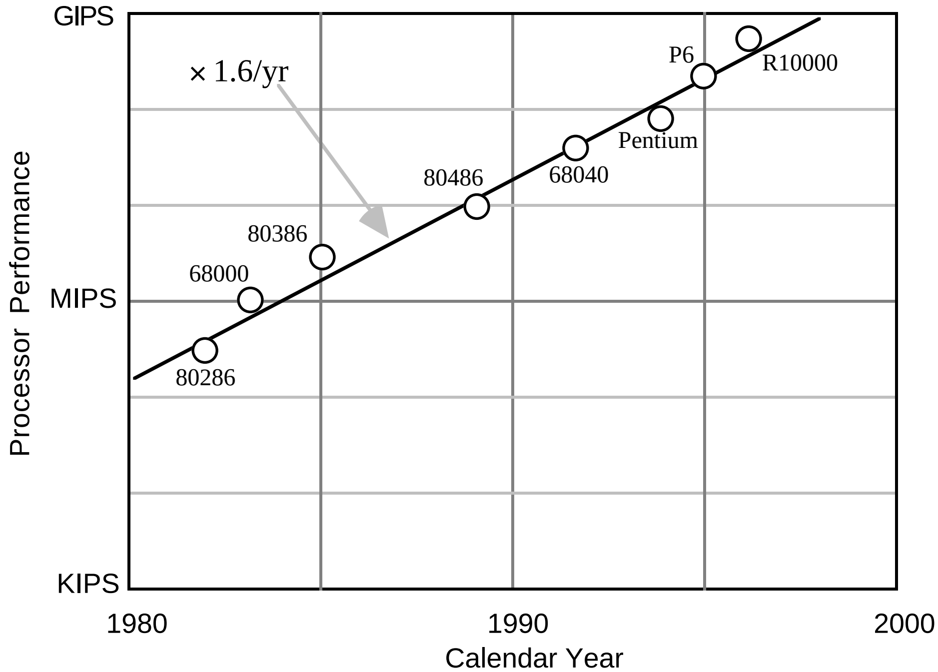

Fig. 1.1. The exponential growth of microprocessor performance, known as Moore’s Law, shown over the past two decades.

Factors contributing to the validity of Moore’s law

Denser circuits
Architectural improvements

Measures of processor performance

Instructions per second (MIPS, GIPS, TIPS, PIPS)
Floating-point operations per second
(MFLOPS, GFLOPS, TFLOPS, PFLOPS)
Running time on benchmark suites

 © Plenum Press, 1999

There is a limit to the speed of a single processor (the speed-of-light argument)

Light travels 30 cm/ns;
signals on wires travel at a fraction of this speed
If signals must travel 1 cm in an instruction cycle,
30 GIPS is the best we can hope for

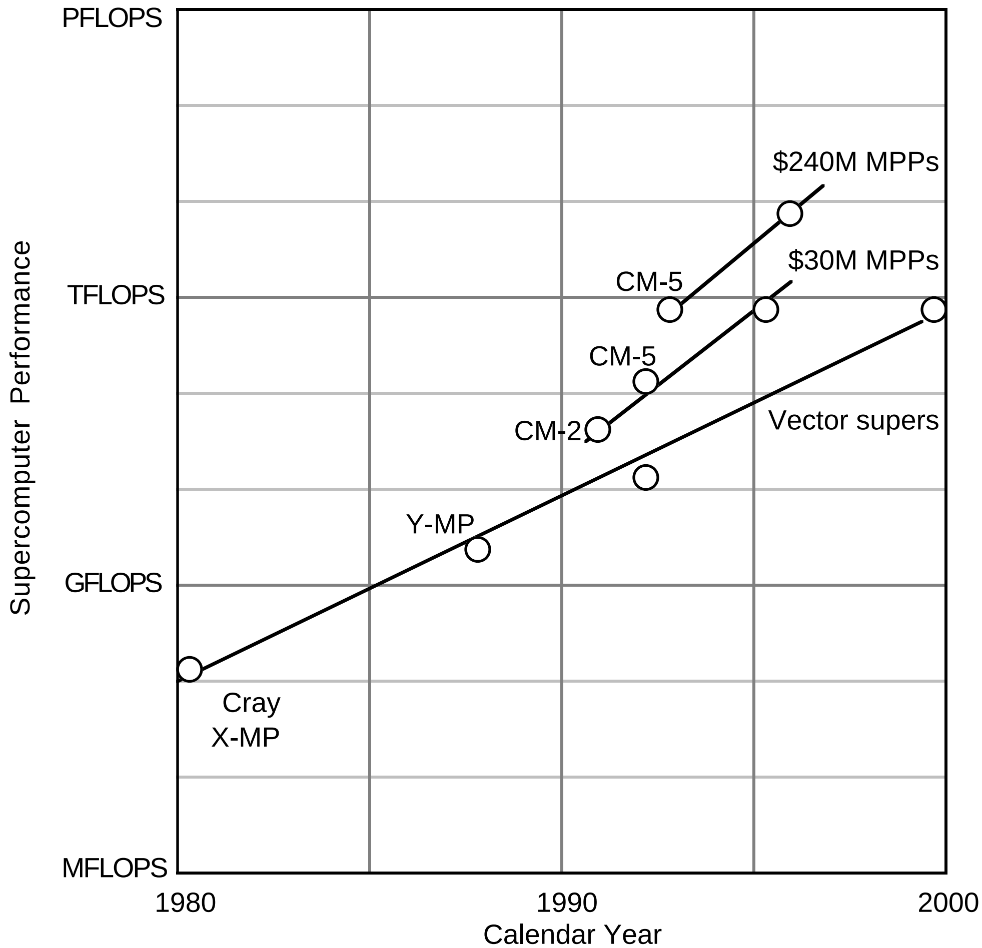

Fig. 1.2. The exponential growth in supercomputer per-formance over the past two decades [Bell92].

 © Plenum Press, 1999

The need for TFLOPS

Modeling of heat transport to the South Pole in the southern oceans [Ocean model: 4096 E-W regions × 1024 N-S regions × 12 layers in depth]

30 000 000 000 FLOP per 10-min iteration ×
300 000 iterations per six-year period =
10^{16} FLOP

Fluid dynamics

1000 × 1000 × 1000 lattice ×
1000 FLOP per lattice point × 10 000 time steps =
10^{16} FLOP

Monte Carlo simulation of nuclear reactor

100 000 000 000 particles to track (for ≈1000 escapes) ×
10 000 FLOP per particle tracked =
10^{15} FLOP

Reasonable running time =
Fraction of hour to several hours (10^3-10^4 s)

Computational power =
10^{16} FLOP / 10^4 s or 10^{15} FLOP / 10^3 s = 10^{12} FLOPS

Why the current quest for PFLOPS?

Same problems, perhaps with finer grids or longer simulated times

 © Plenum Press, 1999

Motivation Summary of Parallel Processing

1. Higher Speed
   Few hours to do 24-hour weather forecasting or tornado warning

2. Higher throughput
   Transaction processing for banks and airlines

3. Higher computational power
   Generate more detailed, accurate and longer simulation e.g. 5 day weather forecasting

- All could be summed up by speedup
- Ideal case to get speedup of *p* with *p* processors
- Actual gain is less than *p* and depends on the architecture and algorithm
- Parallel synergy:
  - Speedup > *p* or ∝ where a task is virtually impossible to perform on a single processor

Status of Computing Power (circa 2000)

GFLOPS on desktop

Apple Macintosh, with G4 processor

TFLOPS in supercomputer center

1152-processor IBM RS/6000 SP

uses a switch-based interconnection network

see IEEE Concurrency, Jan.-Mar. 2000, p. 9

Cray T3E, torus-connected

PFLOPS on drawing board

1M-processor IBM Blue Gene (2005?)

see IEEE Concurrency, Jan.-Mar. 2000, pp. 5-9

32 proc's/chip, 64 chips/board, 8 boards/tower, 64 towers

Processor: 8 threads, on-chip memory, no data cache

Chip: defect-tolerant, row/column rings in a 6 × 6 array

Board: 8 × 8 chip grid organized as 4 × 4 × 4 cube

Tower: Each board linked to 4 neighbors in adjacent towers

System: 32 × 32 × 32 cube of chips, 1.5 MW (water-cooled)

 © Plenum Press, 1999

1.2 A Motivating Example

Sieve of Eratosthenes (ˌer-a-'taas-tha-neez)

for finding all primes in [1, n]

```
 2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
m=2

 2  3     5     7     9    11    13    15    17    19    21    23    25    27    29
   m=3

 2  3     5     7          11    13          17    19          23    25          29
         m=5

 2  3     5     7          11    13          17    19          23                29
               m=7
```

Fig. 1.3. The sieve of Eratosthenes yielding a list of 10 primes for n = 30. Marked elements have been distinguished by erasure from the list.

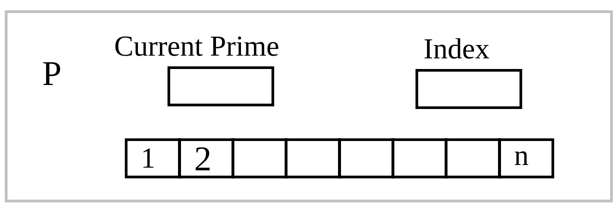

Fig. 1.4. Schematic representation of single-processor solution for the sieve of Eratosthenes.

© Plenum Press, 1999

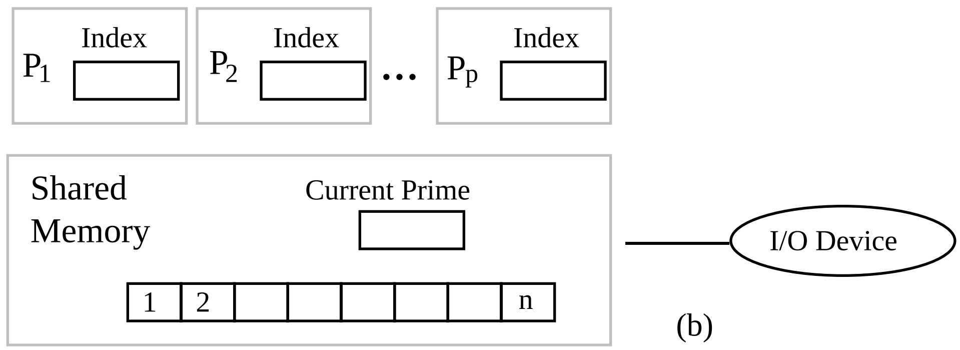

Fig. 1.5. Schematic representation of a control-parallel solution for the sieve of Eratosthenes.

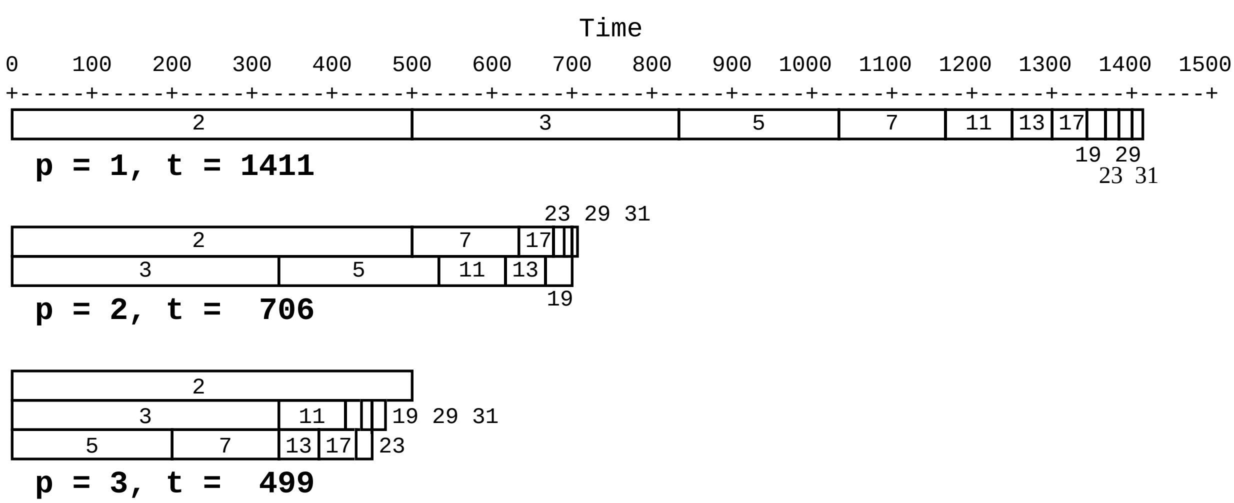

Fig. 1.6. Control-parallel realization of the sieve of Eratosthenes with n = 1000 and 1 ≤ p ≤ 3.

© Plenum Press, 1999

P_1 finds each prime and broadcasts it to all other processors (assume $n/p \le \sqrt{n}$)

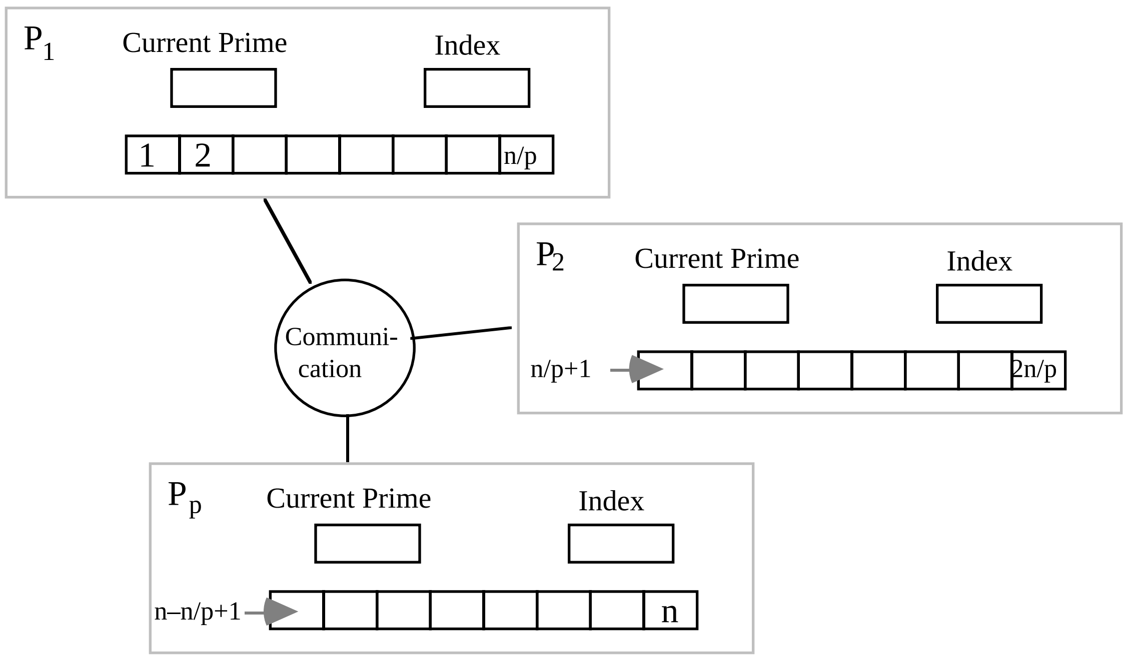

Fig. 1.7. Data-parallel realization of the sieve of Eratosthenes.

 © Plenum Press, 1999

Some reasons for sublinear speed-up

Communication overhead

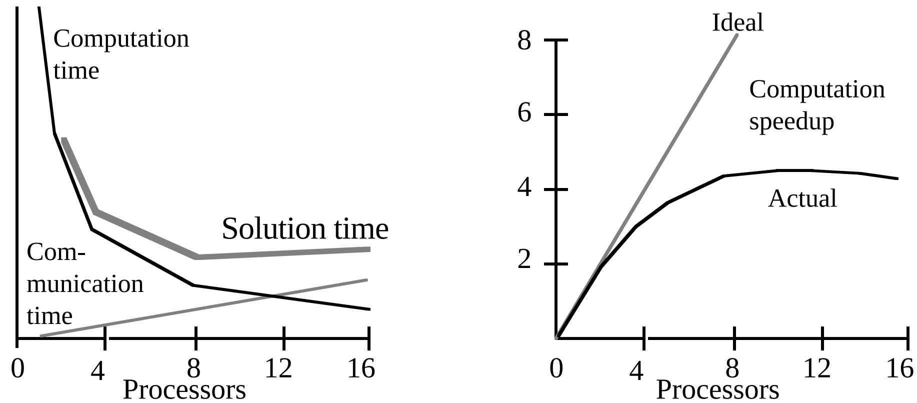

Fig. 1.8. Trade-off between communication time and computation time in the data-parallel realization of the sieve of Eratosthenes.

Input/output overhead

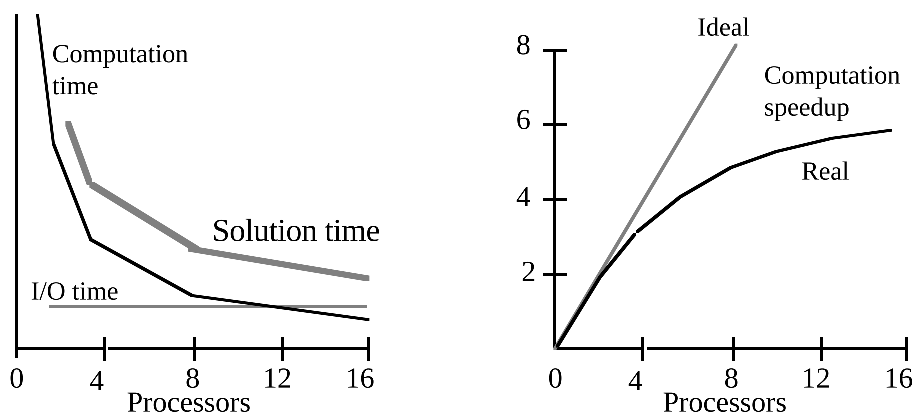

Fig. 1.9. Effect of a constant I/O time on the data-parallel realization of the sieve of Eratosthenes.

 © Plenum Press, 1999

1.3 Parallel Processing Ups and Downs

Early 1900s: 1000s of "computers" (humans + calculators)

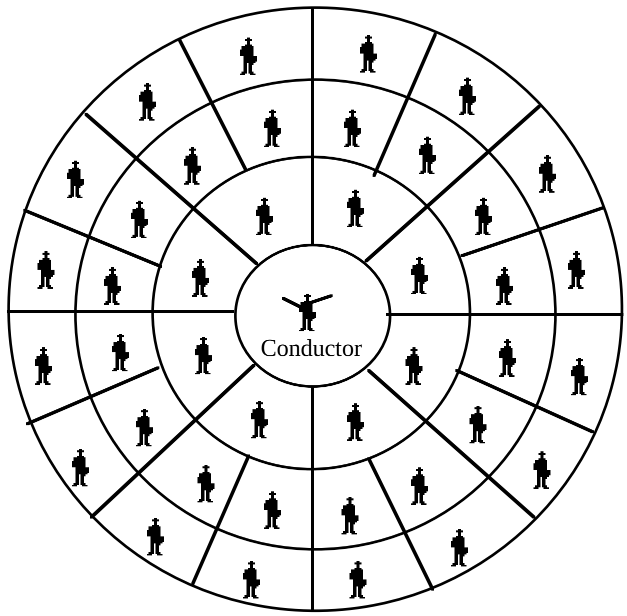

Fig. 1.10. Richardson's circular theater for weather forecasting calculations.

Parallel processing is used in virtually all computers

Compute-I/O overlap, pipelining, multiple function units

But ... in this course we use "parallel processing" in a stricter sense implying the availability of multiple CPUs.

1960s: ILLIAC IV (U Illinois) – 4 quadrants, each 8×8 mesh

1980s: Commercial interest resurfaced; technology was driven by governement contracts. Once funding dried up, many companies went bankrupt.

2000s: The Internet revolution – info providers, multimedia, data mining, etc. need extensive computational power

 © Plenum Press, 1999

1.4 Types of Parallelism: A Taxonomy

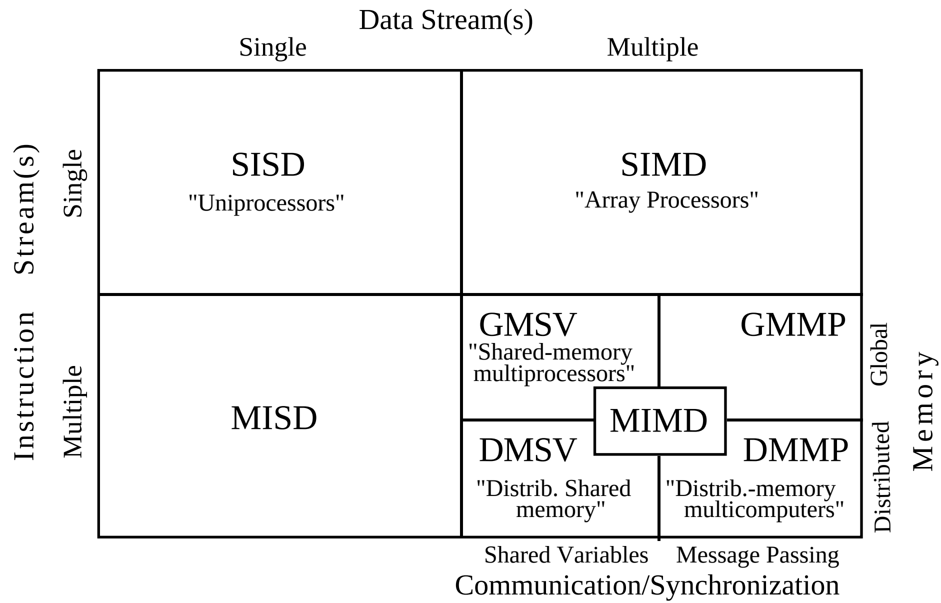

Fig. 1.11. The Flynn-Johnson classification of computer systems.

1.5 Roadblocks to Parallel Processing

a. Grosch's law (economy of scale applies, or computing power is proportional to the square of cost)

b. Minsky's conjecture (speedup is proportional to the logarithm of the number p of processors)

c. Tyranny of IC technology (since hardware becomes about 10 times faster every 5 years, by the time a parallel machine with 10-fold performance is built, uniprocessors will be just as fast)

d. Tyranny of vector supercomputers (vector supercomputers are rapidly improving in performance and offer a familiar programming model and excellent vectorizing compilers; why bother with parallel processors?)

e. The software inertia (Billions of dollars worth of existing software makes it hard to switch to parallel systems)

 © Plenum Press, 1999

f. Amdahl's law

a small fraction f of inherently sequential
or unparallelizable computation
severely limits the speed-up)

$$\text{speedup} \le \frac{1}{f + (1 - f)/p} = \frac{p}{1 + f(p - 1)}$$

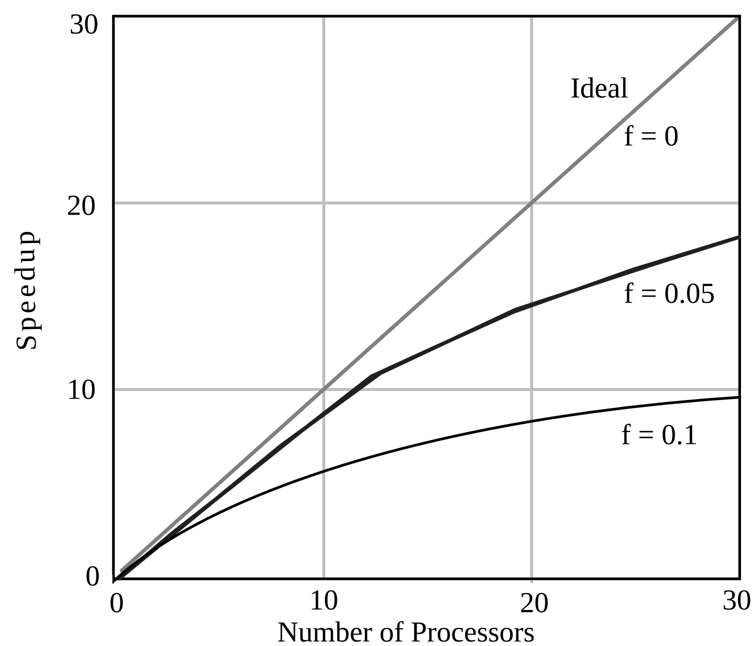

Fig. 1.12. The limit on speed-up according to Amdahl's law.

© Plenum Press, 1999

1.6 Effectiveness of Parallel Processing

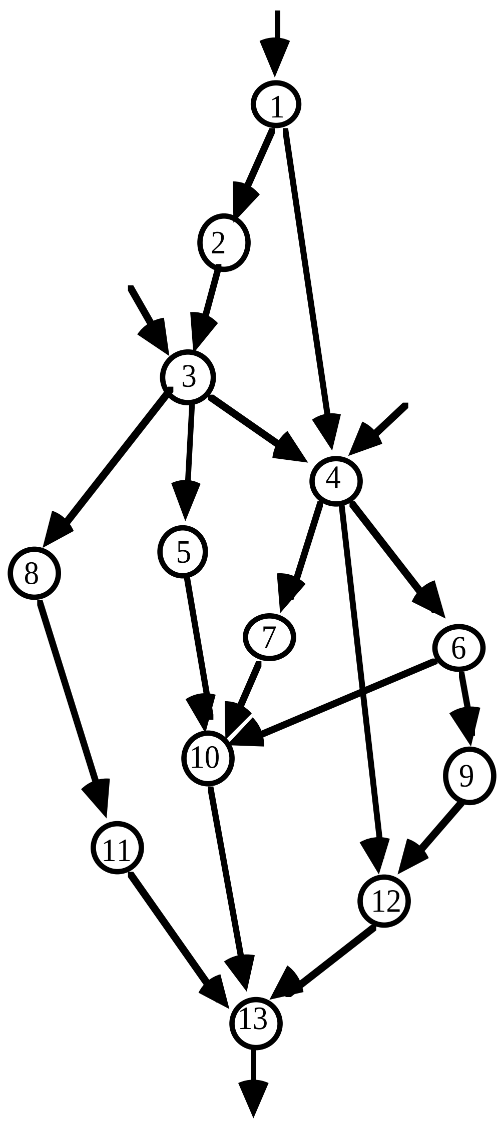

Fig. 1.13. Task graph exhibiting limited inherent parallelism.

© Plenum Press, 1999

Measures used in this course to compare parallel architectures and algorithms [Lee80]:

p	Number of processors
W(p)	Total number of unit operations performed by the p processors; computational work or energy
T(p)	Execution time with p processors; $T(1) = W(1)$ and $T(p) \le W(p)$

$$S(p) \quad \text{Speedup} = \frac{T(1)}{T(p)}$$

$$E(p) \quad \text{Efficiency} = \frac{T(1)}{pT(p)}$$

$$R(p) \quad \text{Redundancy} = \frac{W(p)}{W(1)}$$

$$U(p) \quad \text{Utilization} = \frac{W(p)}{pT(p)}$$

$$Q(p) \quad \text{Quality} = \frac{T^3(1)}{pT^2(p)W(p)}$$

Relationships among the preceding measures:

$$1 \le S(p) \le p \qquad U(p) = R(p)E(p)$$

$$E(p) = \frac{S(p)}{p} \qquad Q(p) = E(p)\frac{S(p)}{R(p)}$$

$$\frac{1}{p} \le E(p) \le U(p) \le 1 \qquad 1 \le R(p) \le \frac{1}{E(p)} \le p$$

$$Q(p) \le S(p) \le p$$

© Plenum Press, 1999

Example: Adding 16 numbers, assuming unit-time additions and ignoring all else, with p = 8

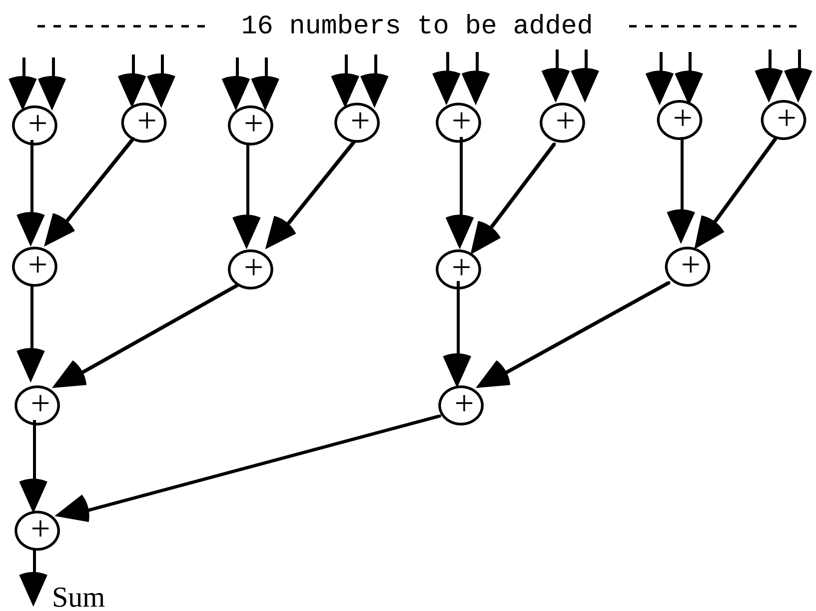

Fig. 1.14. Computation graph for finding the sum of 16 numbers.

Zero-time communication: W(8) = 15 T(8) = 4

E(8) = 15 / (8 × 4) = 47%

S(8) = 15 / 4 = 3.75 R(8) = 15/15 = 1 Q(8) = 1.76

Unit-time communication: W(8) = 22 T(8) = 7

E(8) = 15 / (8 × 7) = 27%

S(8) = 15 / 7 = 2.14 R(8) = 22 / 15 = 1.47 Q(8) = 0.39

 © Plenum Press, 1999