High Performance Computing Systems (CMSC714)

Lecture 13: Isoefficiency and Perf. Modeling

Abhinav Bhatele, Department of Computer Science

UNIVERSITY OF MARYLAND

Announcements

- Project descriptions due tonight
- No lectures next week

Summary of last lecture

- Scalable networks: fat-tree, dragonfly
  - Use high-radix routers
  - Many nodes connected to each switch
- Low network diameter, high bisection bandwidth
- Dynamic routing

Performance analysis methods

- Analytical techniques: use algebraic formulae
  - In terms of data size (n), number of processes (p)
- Time complexity analysis
- Scalability analysis (Isoefficiency)
- Model performance of various operations
  - Analytical models: LogP, alpha-beta model

Isoefficiency

- Relationship between problem size and number of processors to maintain a certain level of efficiency
- At what rate should we increase problem size with respect to number of processors to keep efficiency constant

Speedup and efficiency

- Speedup: Ratio of execution time on one process to that on p processes

$$\text{Speedup} = \frac{t_1}{t_p}$$

- Efficiency: Speedup per process

$$\text{Efficiency} = \frac{t_1}{t_p \times p}$$

Efficiency in terms of overhead

- Total time spent in all processes = (useful) computation + overhead (extra computation + communication + idle time)

$$p \times t_p = t_1 + t_o$$

$$\text{Efficiency} = \frac{t_1}{t_p \times p} = \frac{t_1}{t_1 + t_o} = \frac{1}{1 + \frac{t_o}{t_1}}$$

Isoefficiency function

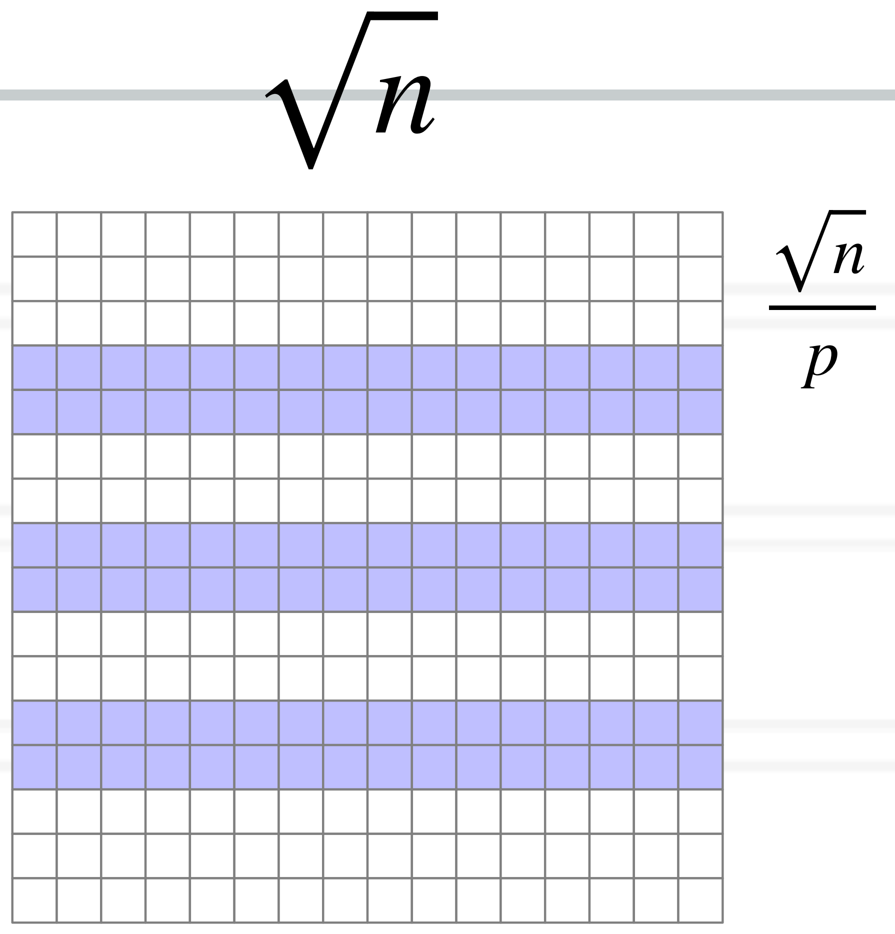
$$\text{Efficiency} = \frac{1}{1 + \frac{t_o}{t_1}}$$

- Efficiency is constant if t_o / t_l is constant (K)

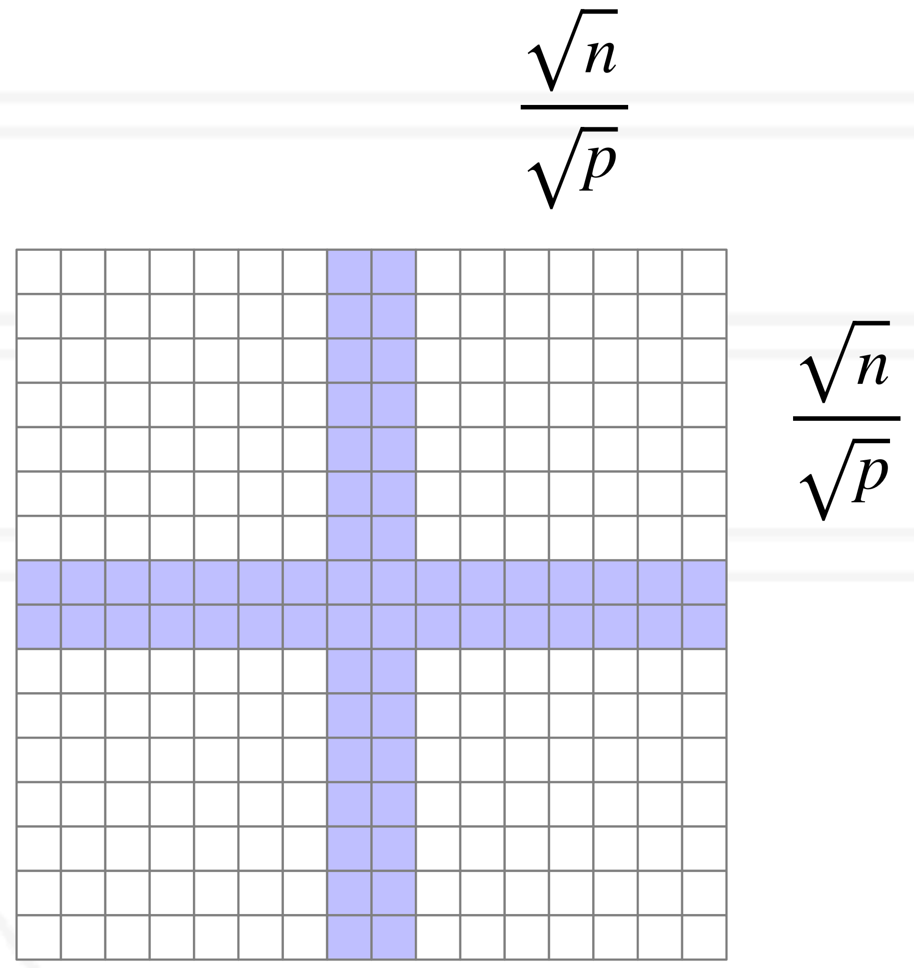
$$t_o = K \times t_1$$

Isoefficiency analysis

- 1D decomposition:
  - Computation:
  - Communication:

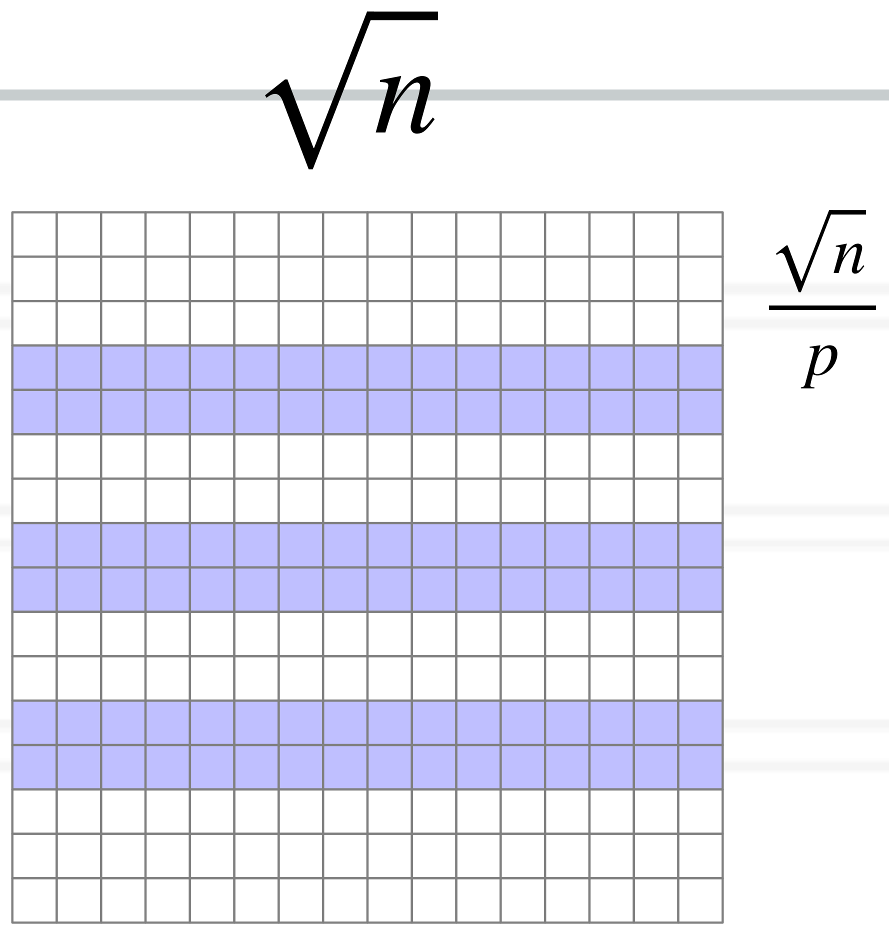


- 2D decomposition:
  - Computation:
  - Communication

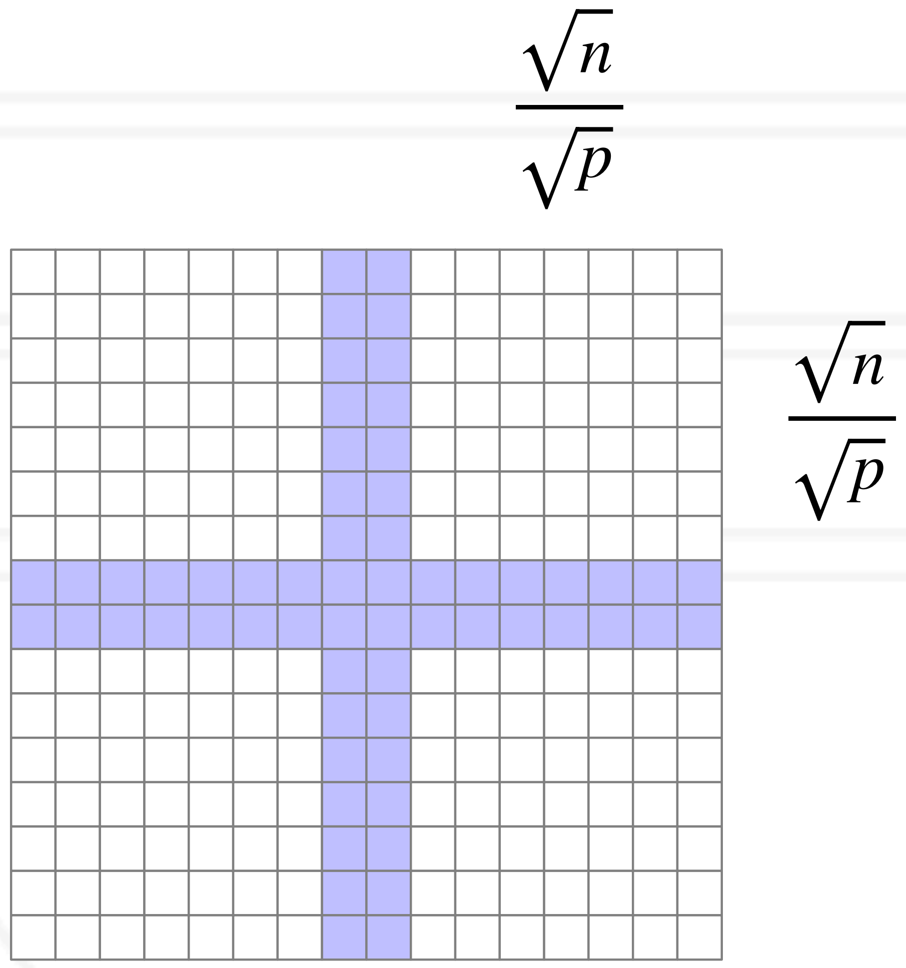


Isoefficiency analysis

- 1D decomposition:
  - Computation: $\sqrt{n} \times \frac{\sqrt{n}}{p} = \frac{n}{p}$
  - Communication:

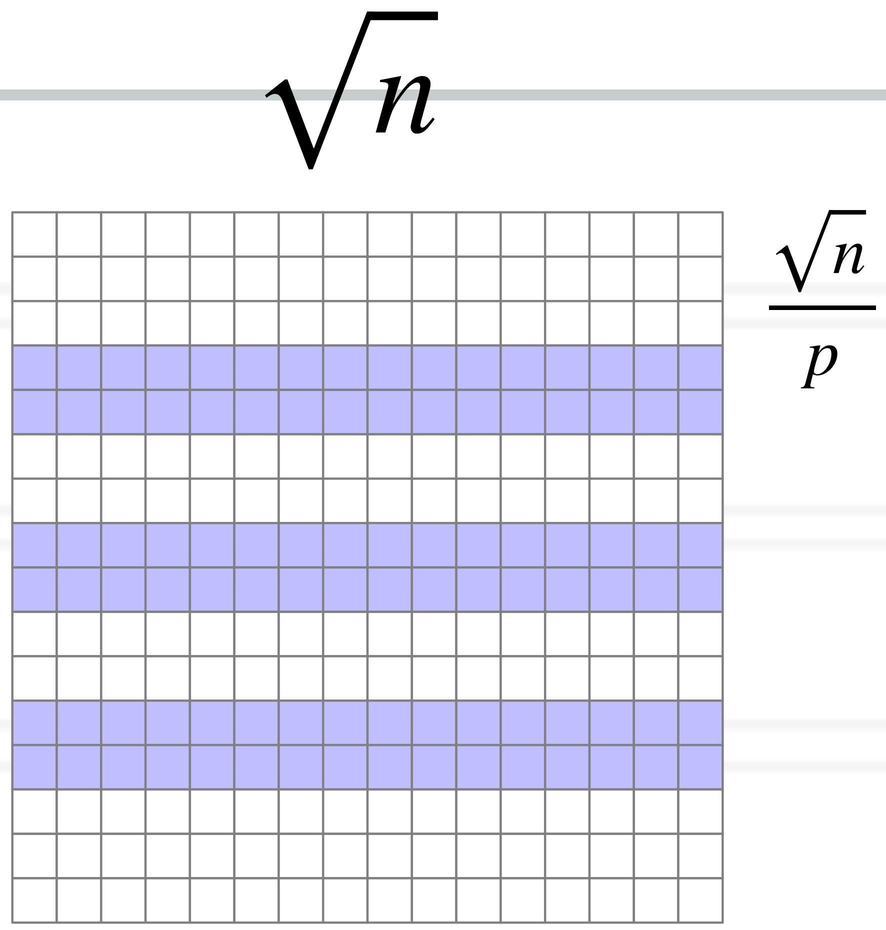


- 2D decomposition:
  - Computation:
  - Communication

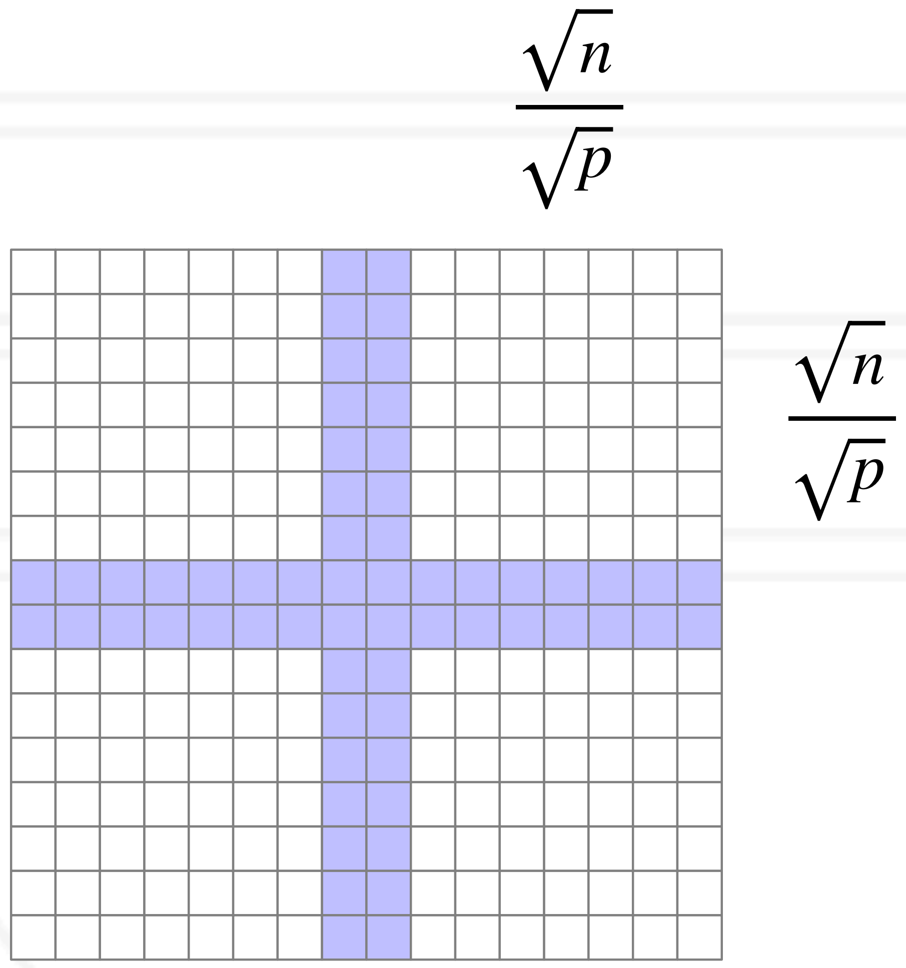


Isoefficiency analysis

- 1D decomposition:
  - Computation: $\sqrt{n} \times \frac{\sqrt{n}}{p} = \frac{n}{p}$
  - Communication: $2 \times \sqrt{n}$



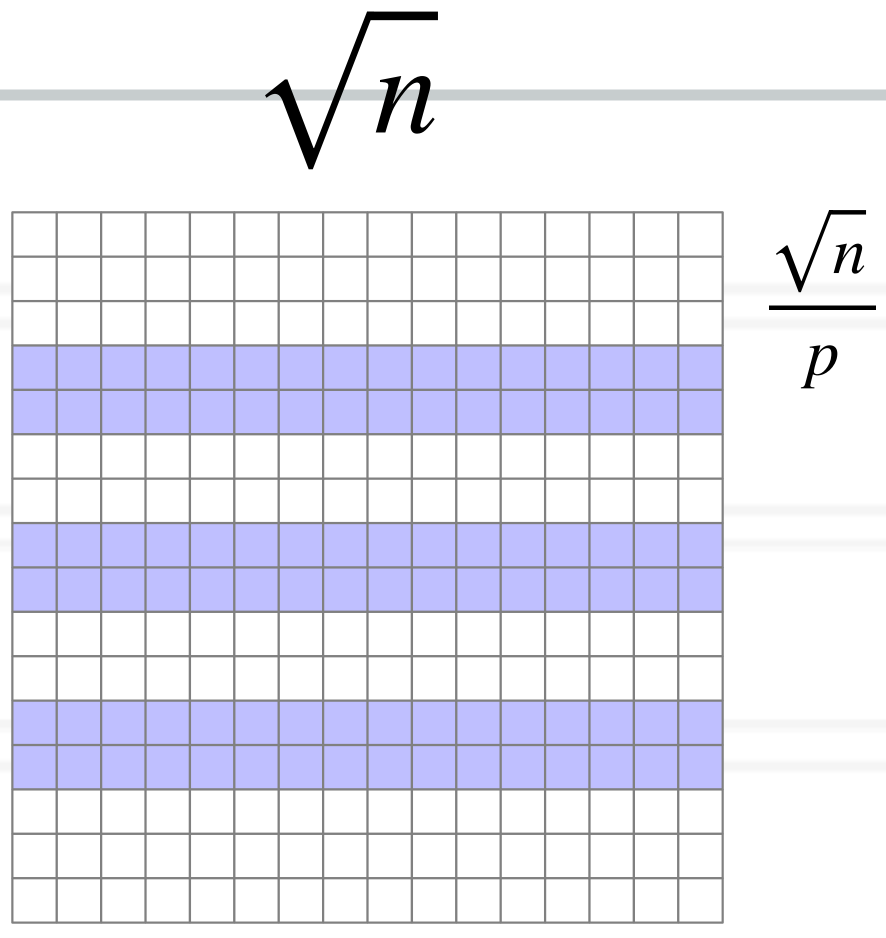
- 2D decomposition:
  - Computation:
  - Communication



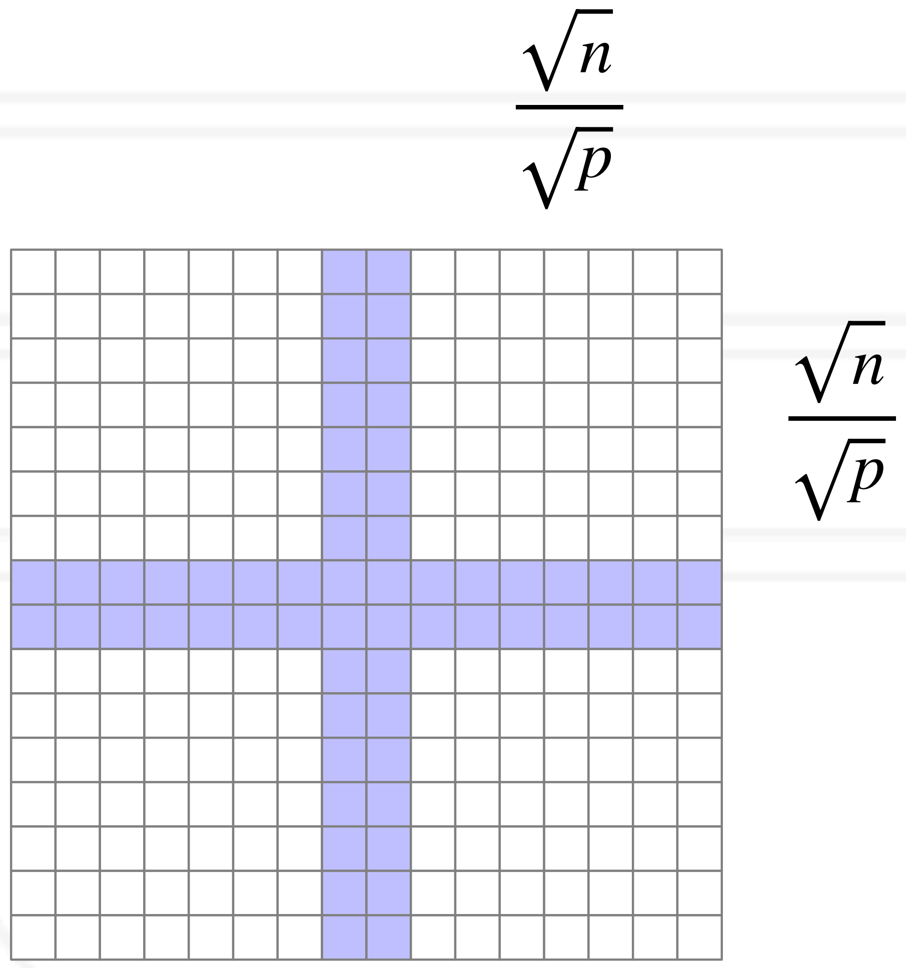
Isoefficiency analysis

- 1D decomposition:
  - Computation: $\sqrt{n} \times \frac{\sqrt{n}}{p} = \frac{n}{p}$
  - Communication: $2 \times \sqrt{n}$

$$\frac{t_0}{t_1} = \frac{2 \times \sqrt{n}}{\frac{n}{p}} = \frac{2 \times p}{\sqrt{n}}$$



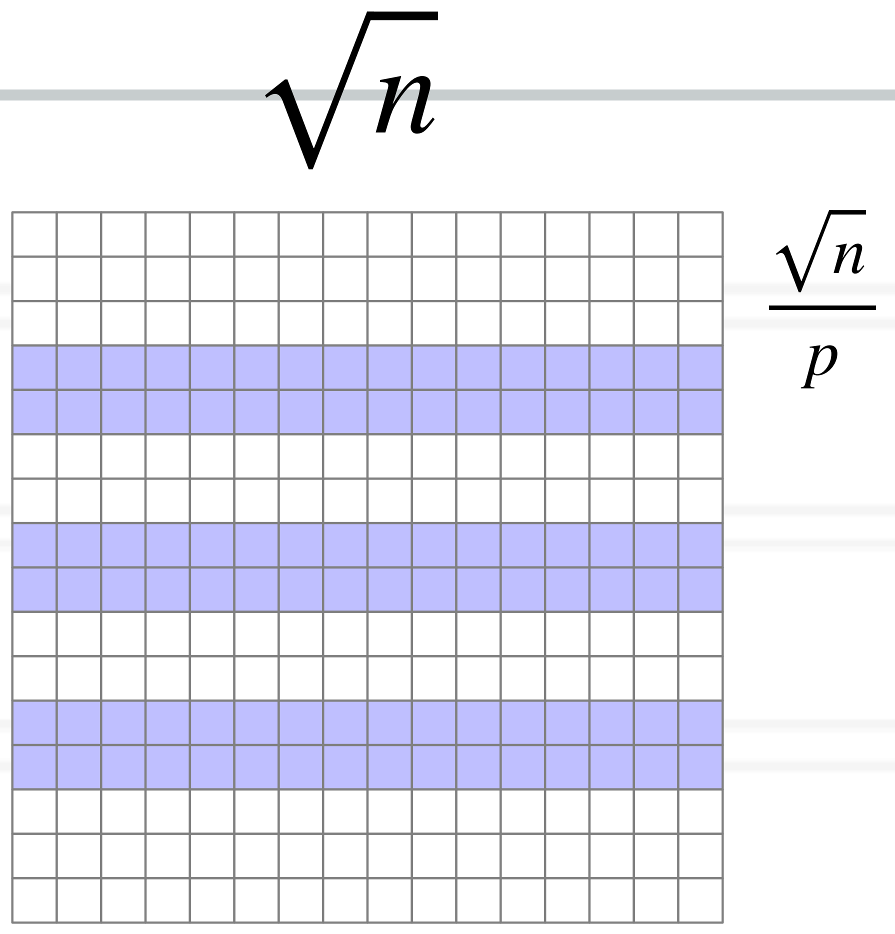
- 2D decomposition:
  - Computation:
  - Communication



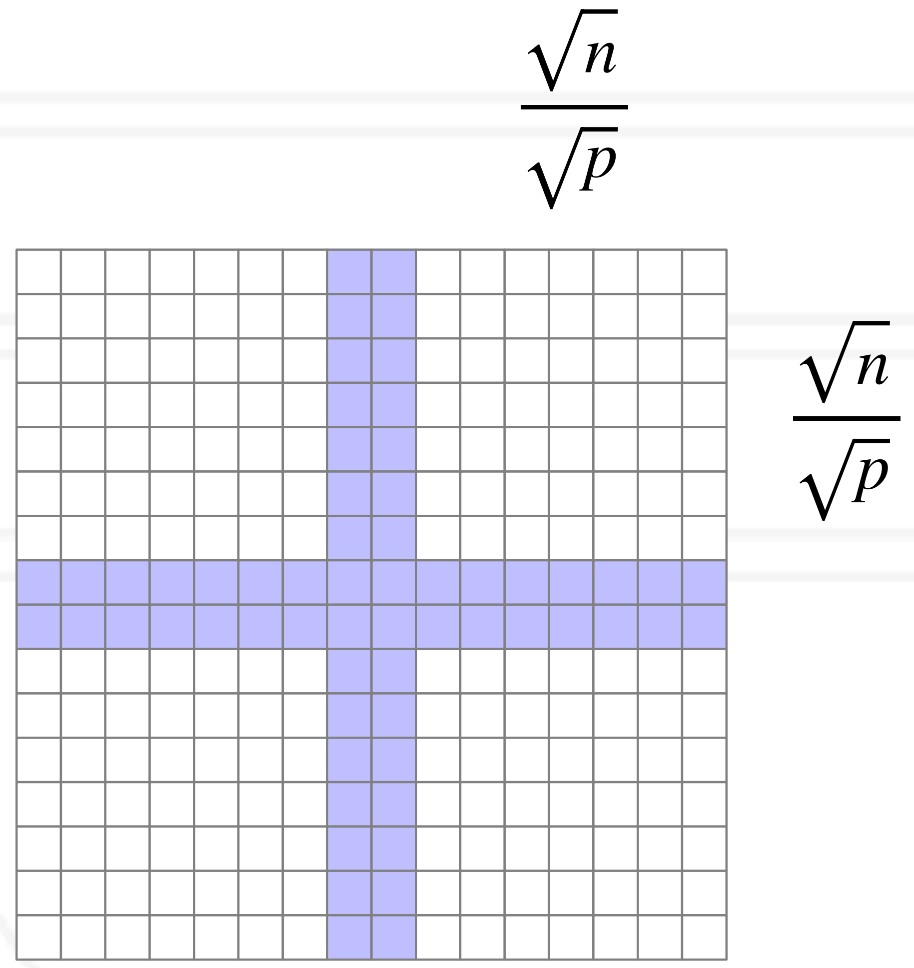
Isoefficiency analysis

- 1D decomposition:
  - Computation: $\sqrt{n} \times \frac{\sqrt{n}}{p} = \frac{n}{p}$
  - Communication: $2 \times \sqrt{n}$

$$\frac{t_0}{t_1} = \frac{2 \times \sqrt{n}}{\frac{n}{p}} = \frac{2 \times p}{\sqrt{n}}$$



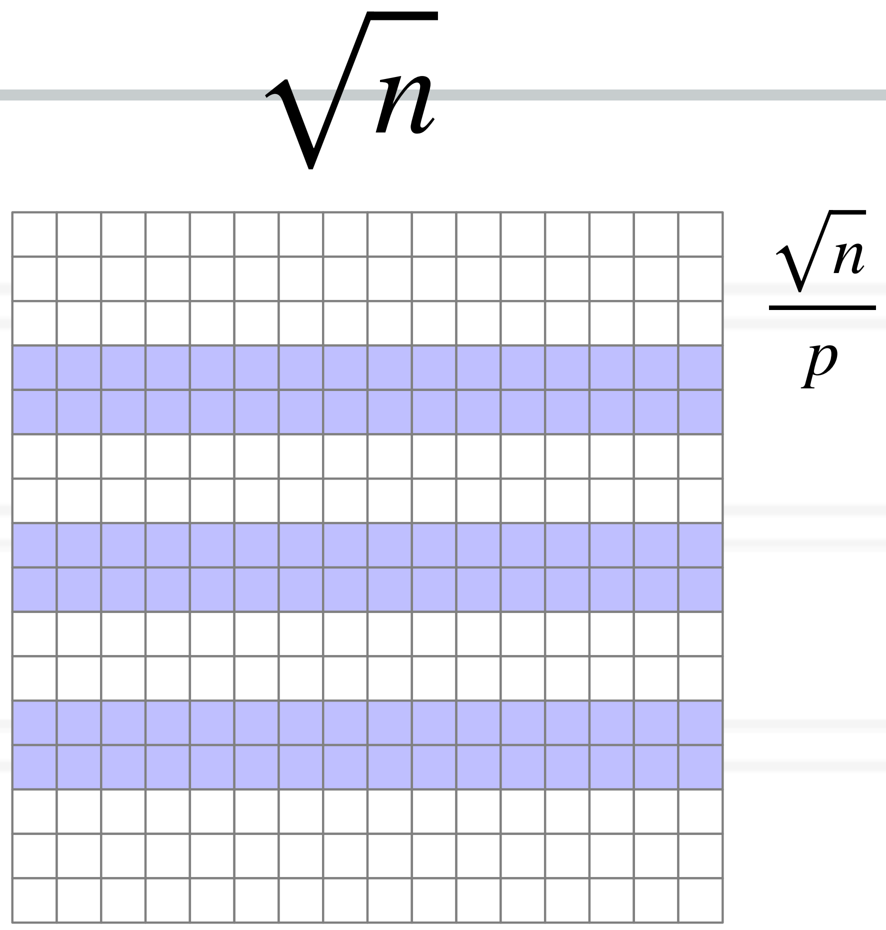
- 2D decomposition:
  - Computation: $\frac{\sqrt{n}}{\sqrt{p}} \times \frac{\sqrt{n}}{\sqrt{p}} = \frac{n}{p}$
  - Communication



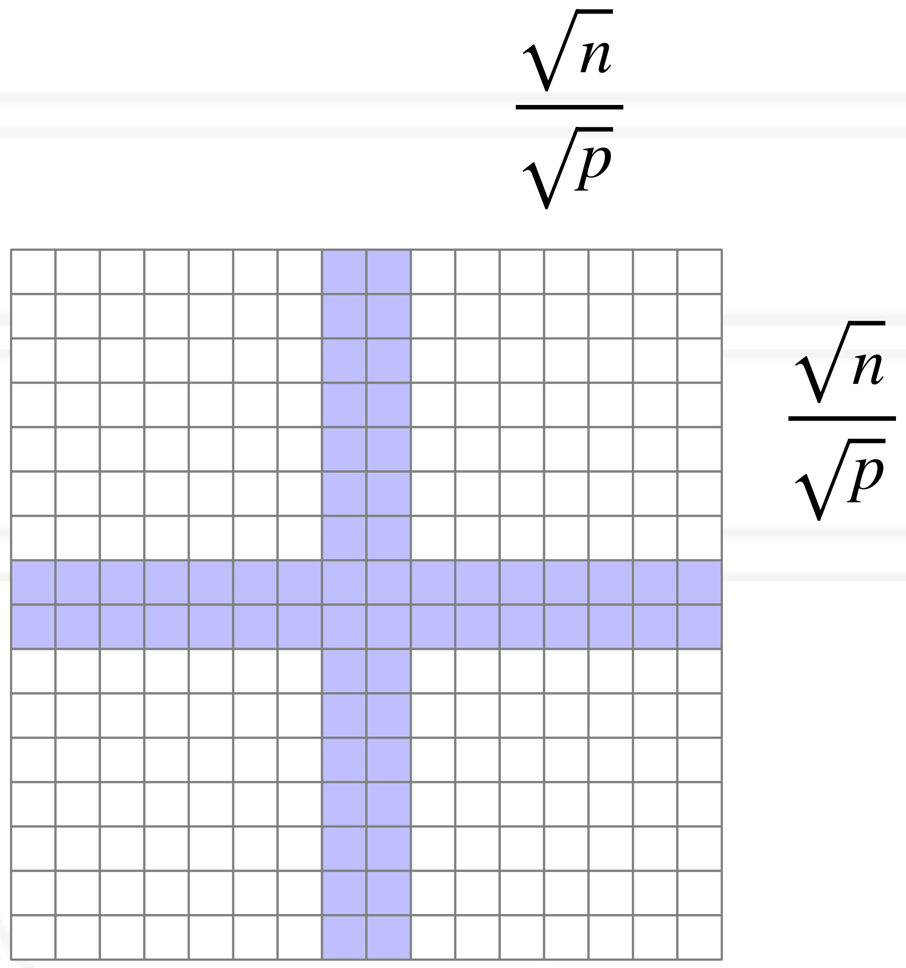
Isoefficiency analysis

- 1D decomposition:
  - Computation: $\sqrt{n} \times \frac{\sqrt{n}}{p} = \frac{n}{p}$
  - Communication: $2 \times \sqrt{n}$

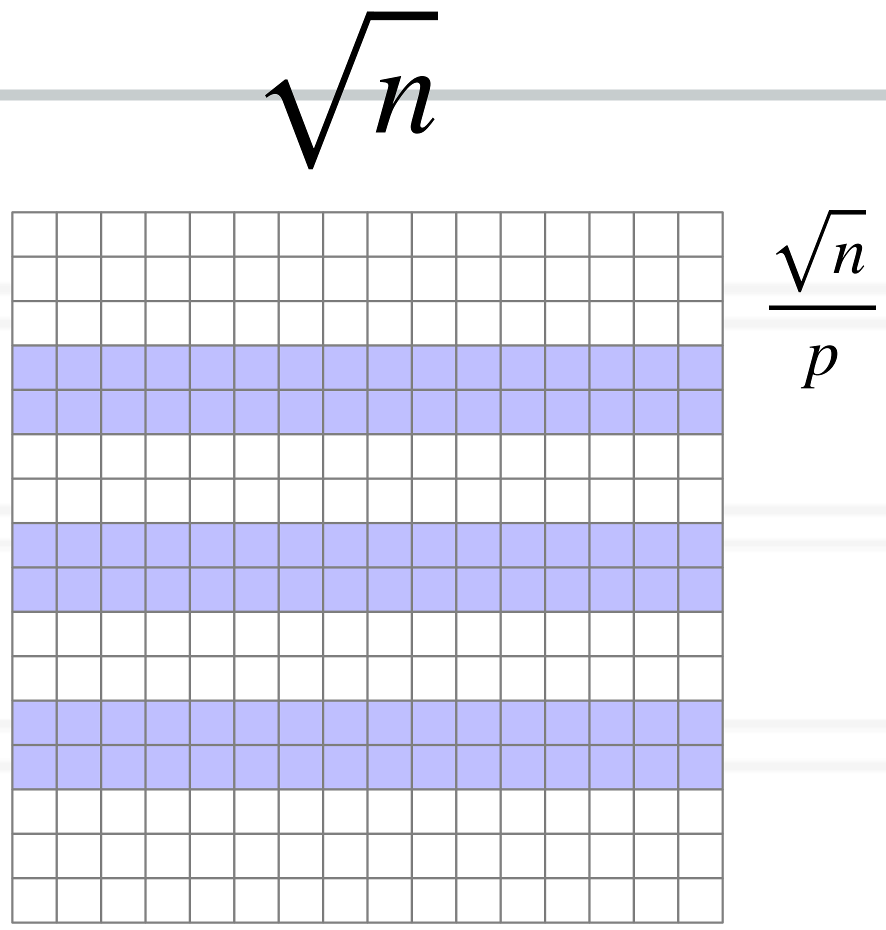
$$\frac{t_0}{t_1} = \frac{2 \times \sqrt{n}}{\frac{n}{p}} = \frac{2 \times p}{\sqrt{n}}$$



- 2D decomposition:
  - Computation: $\frac{\sqrt{n}}{\sqrt{p}} \times \frac{\sqrt{n}}{\sqrt{p}} = \frac{n}{p}$
  - Communication $4 \times \frac{\sqrt{n}}{\sqrt{p}}$

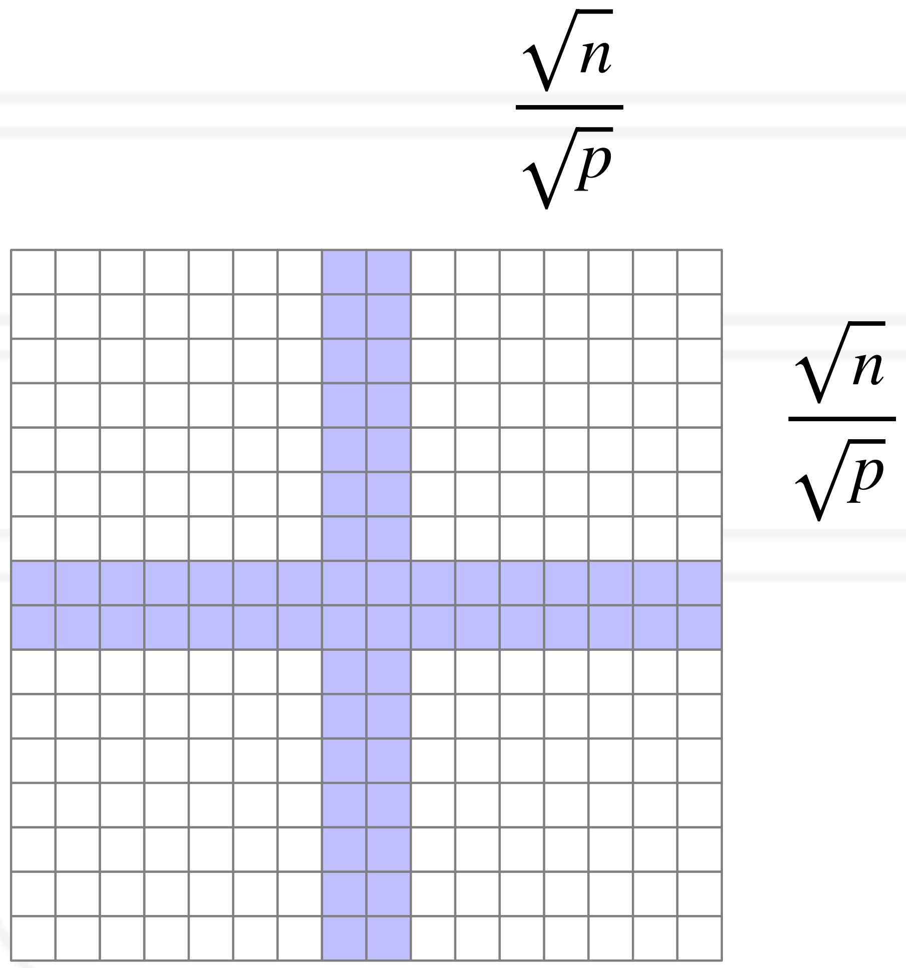


Isoefficiency analysis



- 1D decomposition:
  - Computation: $\sqrt{n} \times \frac{\sqrt{n}}{p} = \frac{n}{p}$
  - Communication: $2 \times \sqrt{n}$

$$\frac{t_0}{t_1} = \frac{2 \times \sqrt{n}}{\frac{n}{p}} = \frac{2 \times p}{\sqrt{n}}$$



- 2D decomposition:
  - Computation: $\frac{\sqrt{n}}{\sqrt{p}} \times \frac{\sqrt{n}}{\sqrt{p}} = \frac{n}{p}$
  - Communication $4 \times \frac{\sqrt{n}}{\sqrt{p}}$

$$\frac{t_0}{t_1} = \frac{4 \times \frac{\sqrt{n}}{\sqrt{p}}}{\frac{n}{p}} = \frac{4 \times \sqrt{p}}{\sqrt{n}}$$

Performance Modeling

- Model the performance of a parallel application
- Different methods
  - Analytical
  - Empirical
  - Simulation

LogP model

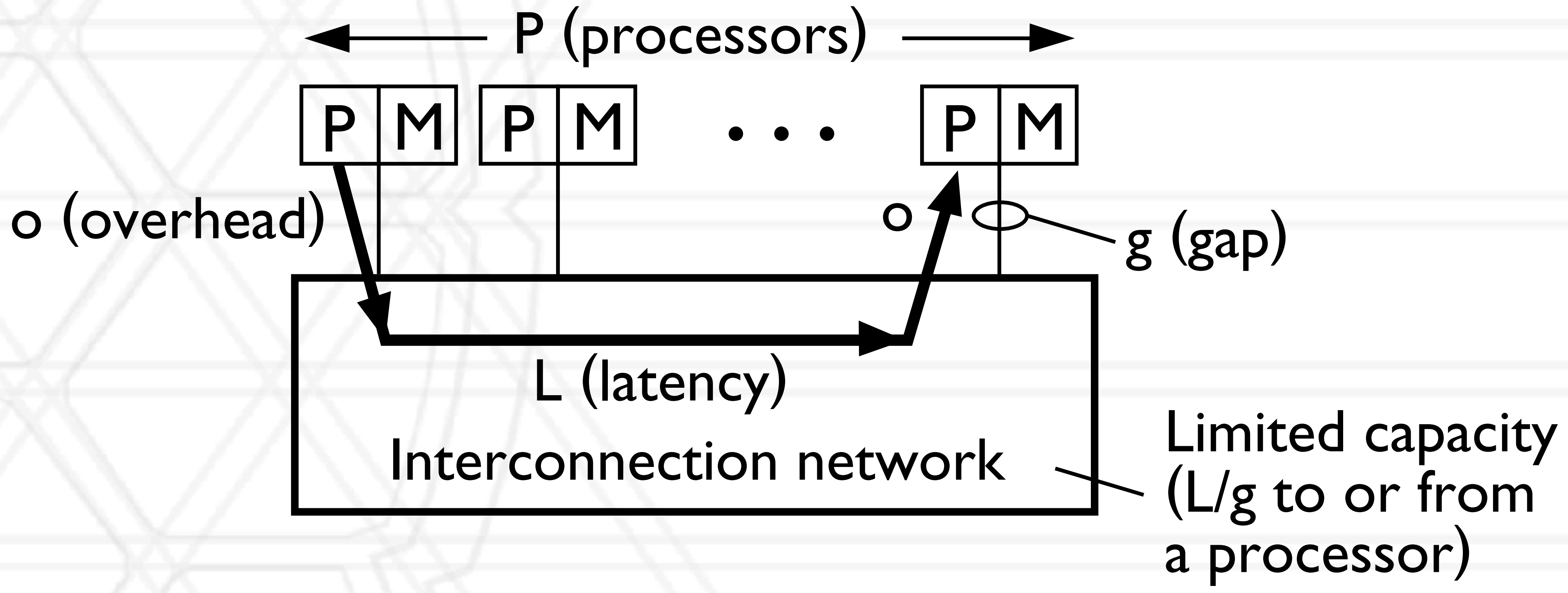
- Model for communication on an interconnection network

L: latency or delay

O: overhead (processor busy in communication)

g: gap

P: number of processors / processes



1/g = bandwidth

alpha + n * beta model

- Another model for communication

$$T_{\text{comm}} = \alpha + n \times \beta$$

α: latency

n: size of message

β: bandwidth

Questions?

UNIVERSITY OF MARYLAND

Abhinav Bhatele

5218 Brendan Iribe Center (IRB) / College Park, MD 20742

phone: 301.405.4507 / e-mail: bhatele@cs.umd.edu