

The impact of architecture design on software

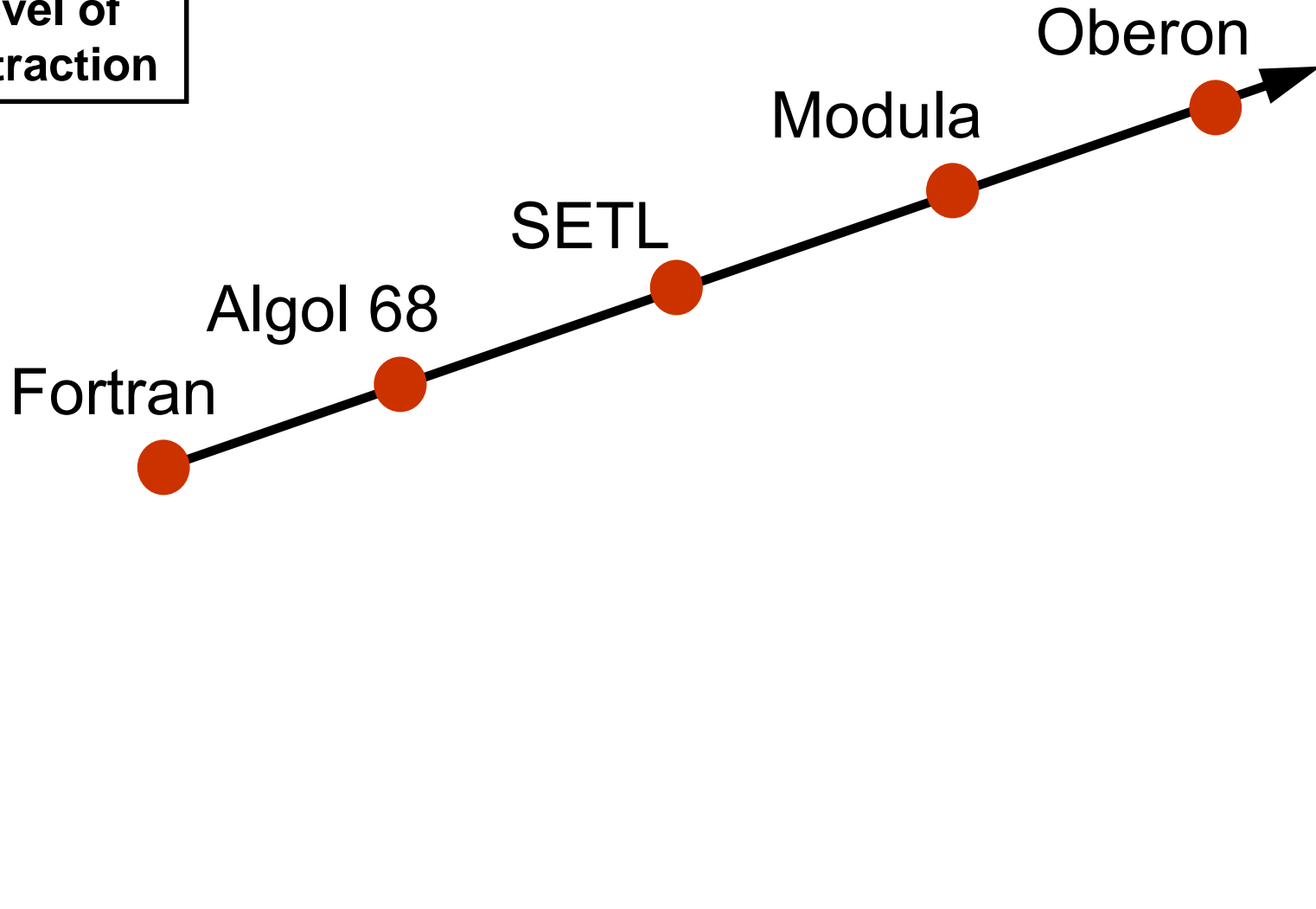
John van Rosendale
Department of Energy

Issues

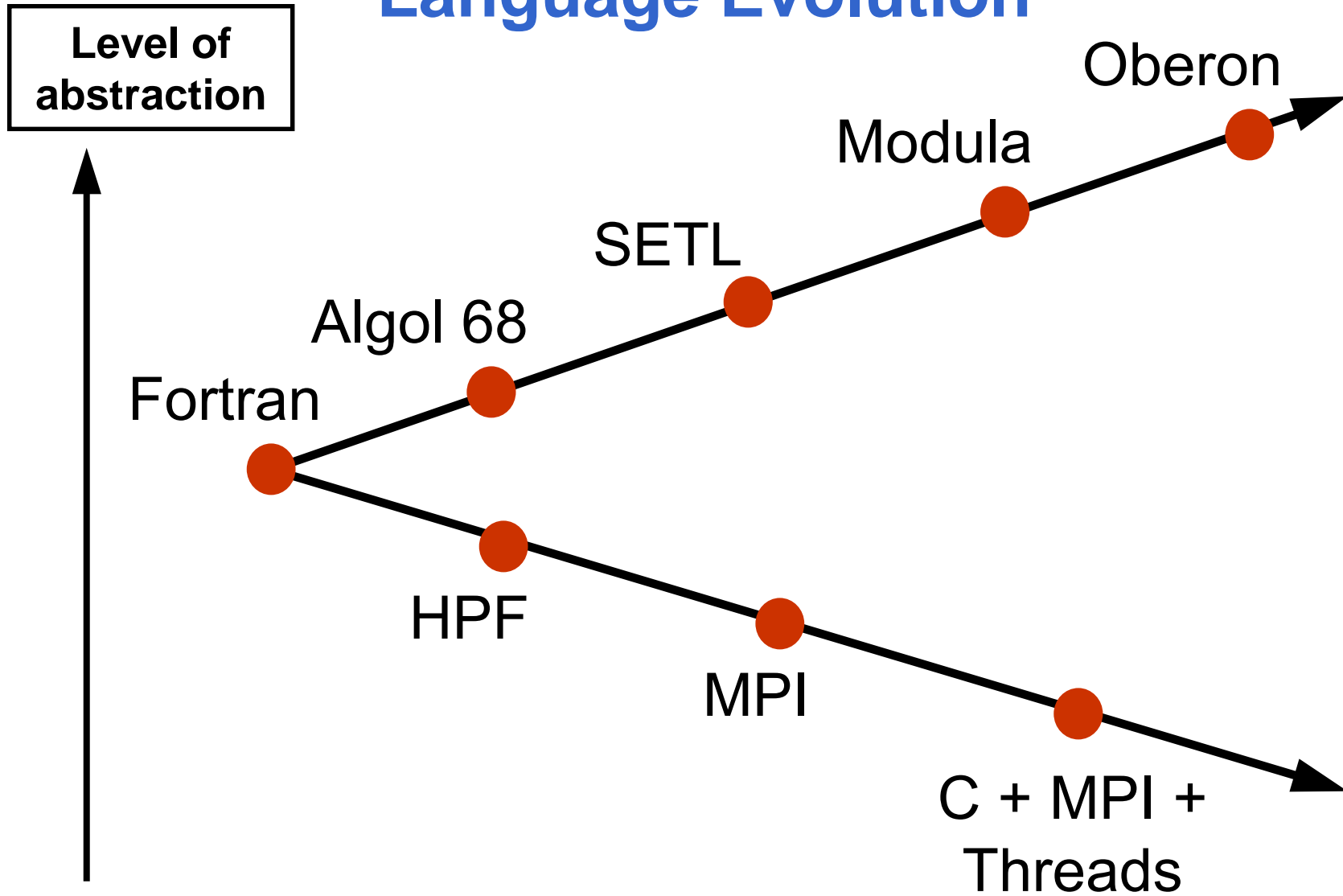
- Programming models
- How much parallelism is needed?
- The impact of architectural choices on software

Language Evolution

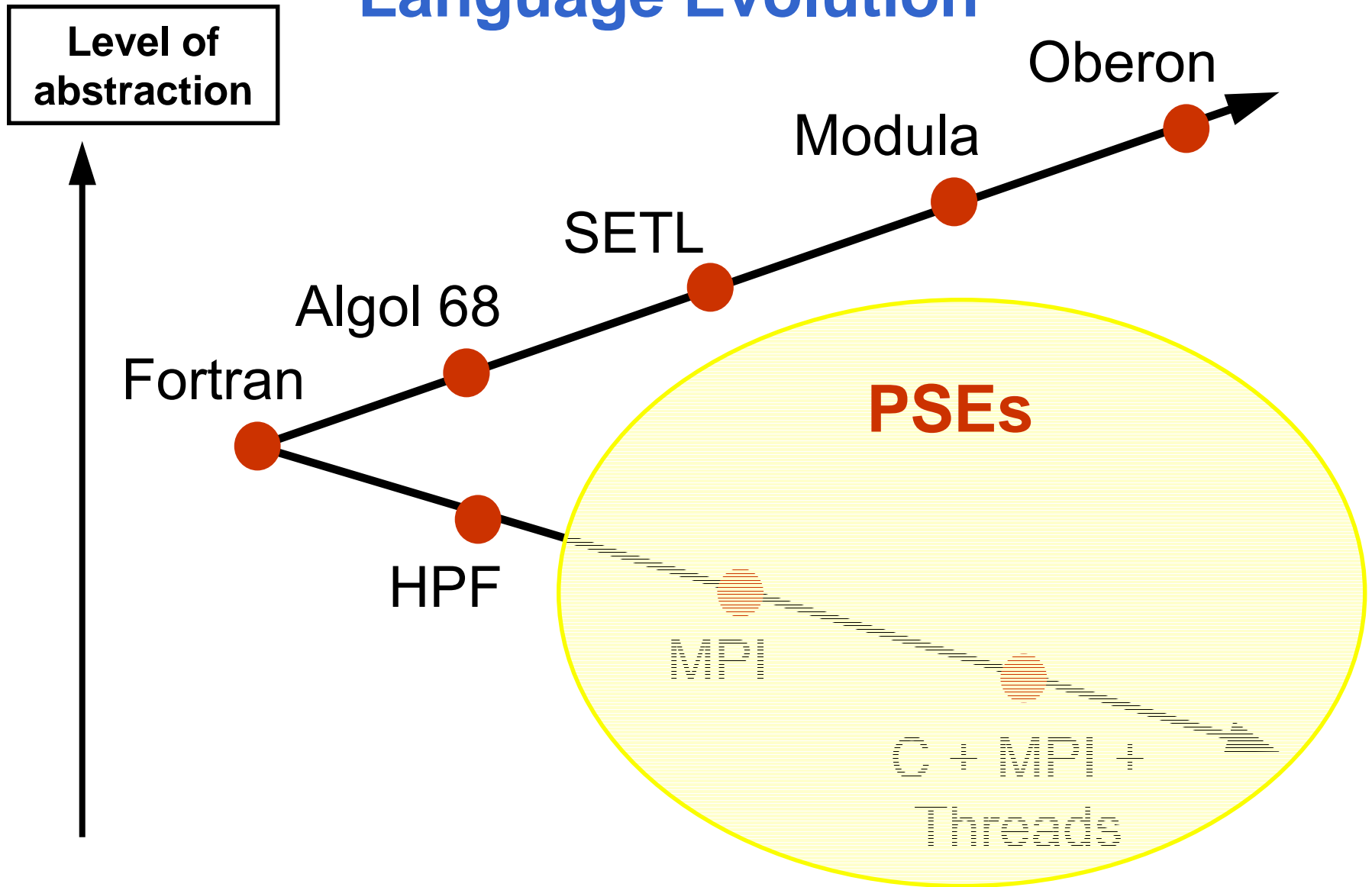
Level of
abstraction



Language Evolution



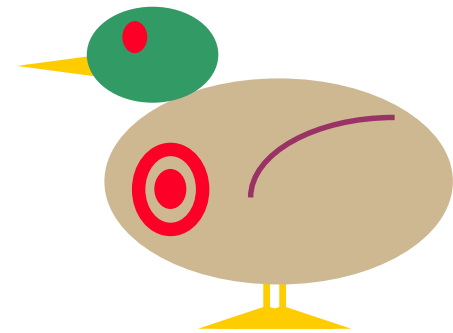
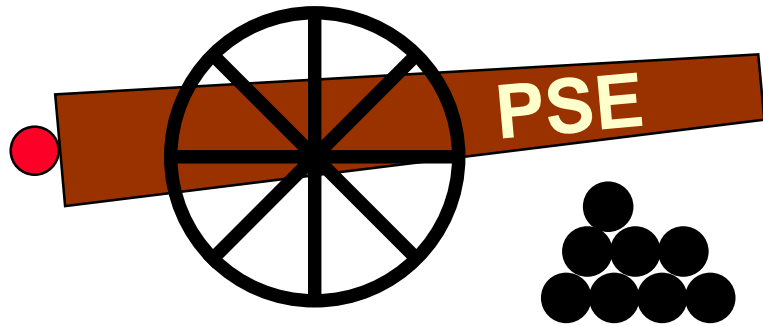
Language Evolution



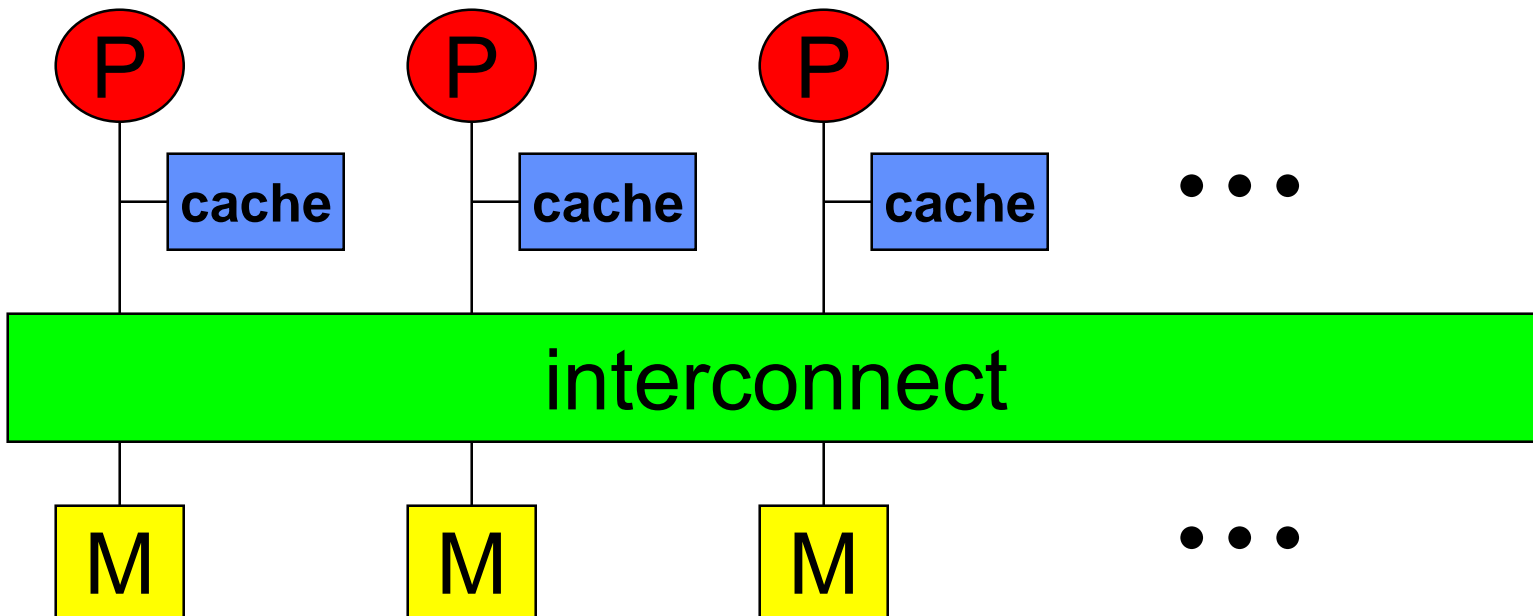
Programming models

- Current parallel scientific programming environments are driven almost entirely by performance issues
- Reasons for this surely include:
 - the relatively modest commercial impact of parallel computing
 - continued evolution of parallel hardware
- PSEs continue to improve, and to some degree, obviate the need for effective high-level parallel languages

Sitting duck metaphor for PSE-utility



How much parallelism?



Parallelism(cont.)

Let

T = main memory latency

R = cache hit ratio

P = intrinsic parallelism of architecture

(average # of concurrently executing threads)

Then the architecture speed in flops/ops satisfies:

$$S \leq P / (1-R) \cdot T$$

Parallelism(cont.)

So for a petaflop architecture with

Cache hit ratio = 90%

Memory latency = $1\mu\text{s}$

$$\begin{aligned} P &\geq (1-R) \cdot T \cdot S \\ &= 0.1 \cdot 1\mu\text{s} \cdot 10^{15} \text{ops/s} \\ &= 10^8 \end{aligned}$$

That is, 100 million-way parallelism is a minimum

Architecture Choices

Price / performance



Beowulf-style
architecture

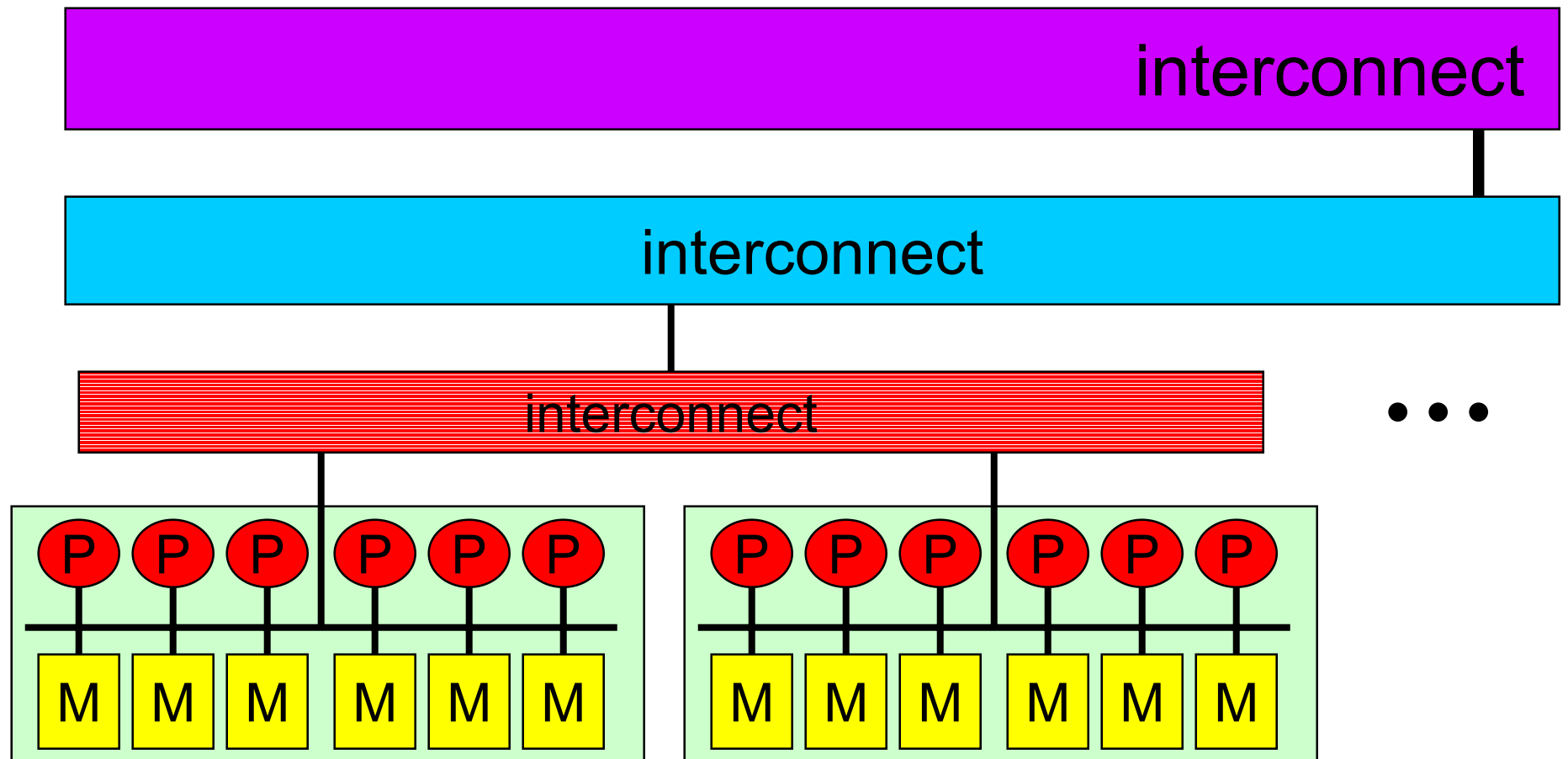
COTs processors
+ custom
interconnect

Totally custom
design



software / application
simplicity

Hierarchical cluster architecture



Programming complexity increases with:

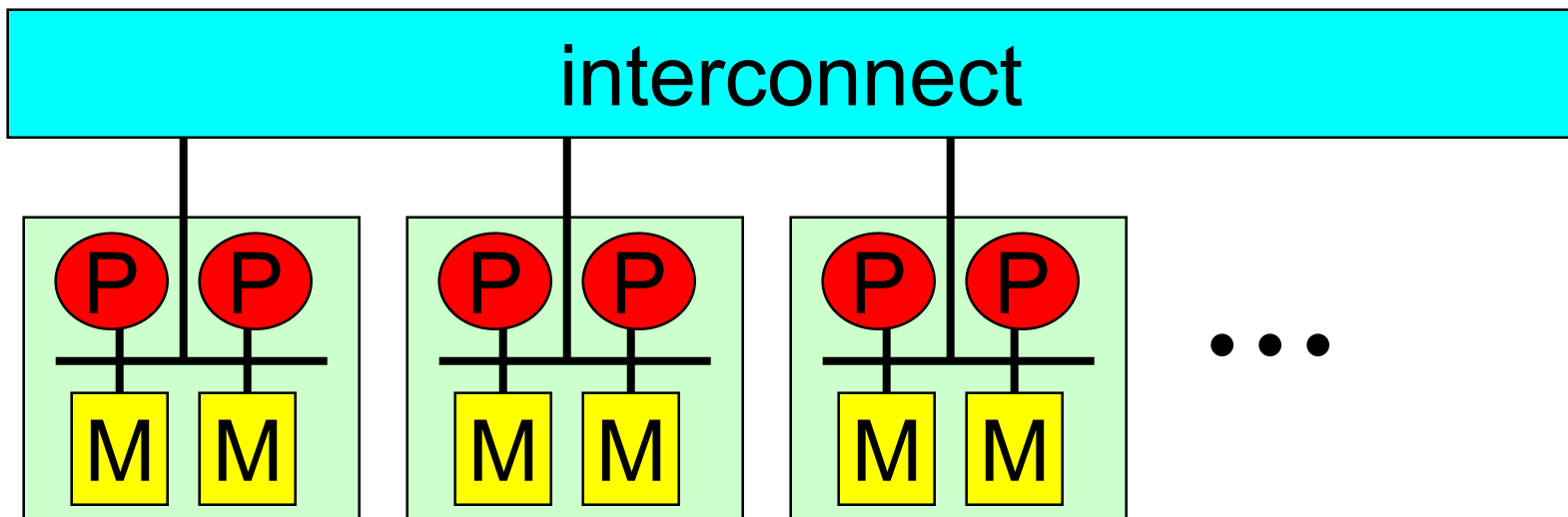
- Number of levels
- Performance differences between levels
- Semantic variation between levels

But:

- Market forces will drive us to deep hierarchies
- Manufacturers will pay most attention to the bottom level
- Bizantine semantics is likely

PIM-based Paragon-style architecture

- “Flat” MPI programming model
- Reduced memory latency implies reduced parallelism



Observations

- It will take an amazing amount of parallelism to saturate foreseeable petaflop architectures
- The parallelism required is strongly architecture dependent
- Architectures supporting reasonable programming models and hence a wide spectrum of applications are critical

Observations (cont.)

- New architectures may open up the possibility of more effective parallel programming models.
- We should continue to explore the idea of effective high-level parallel languages.