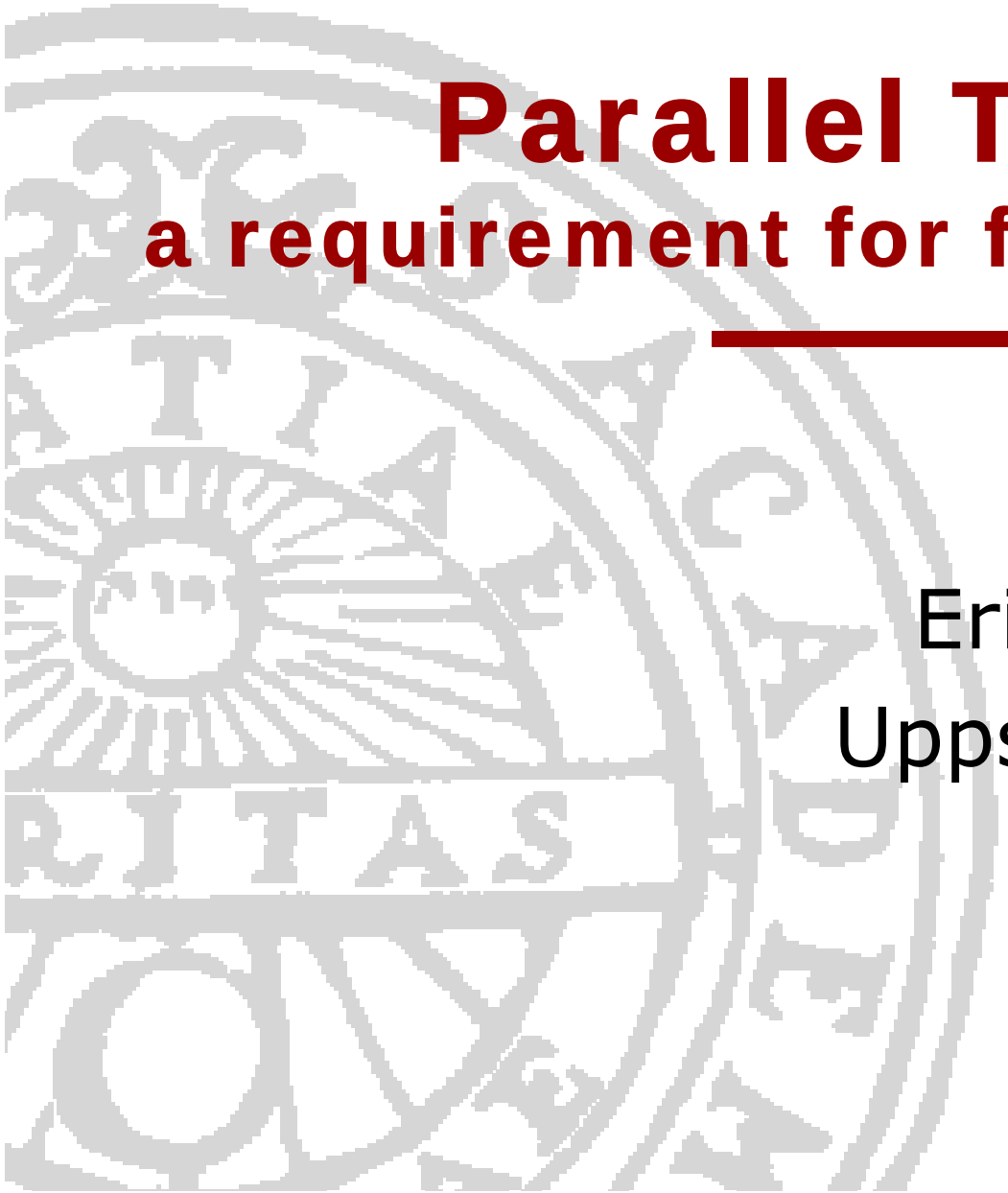
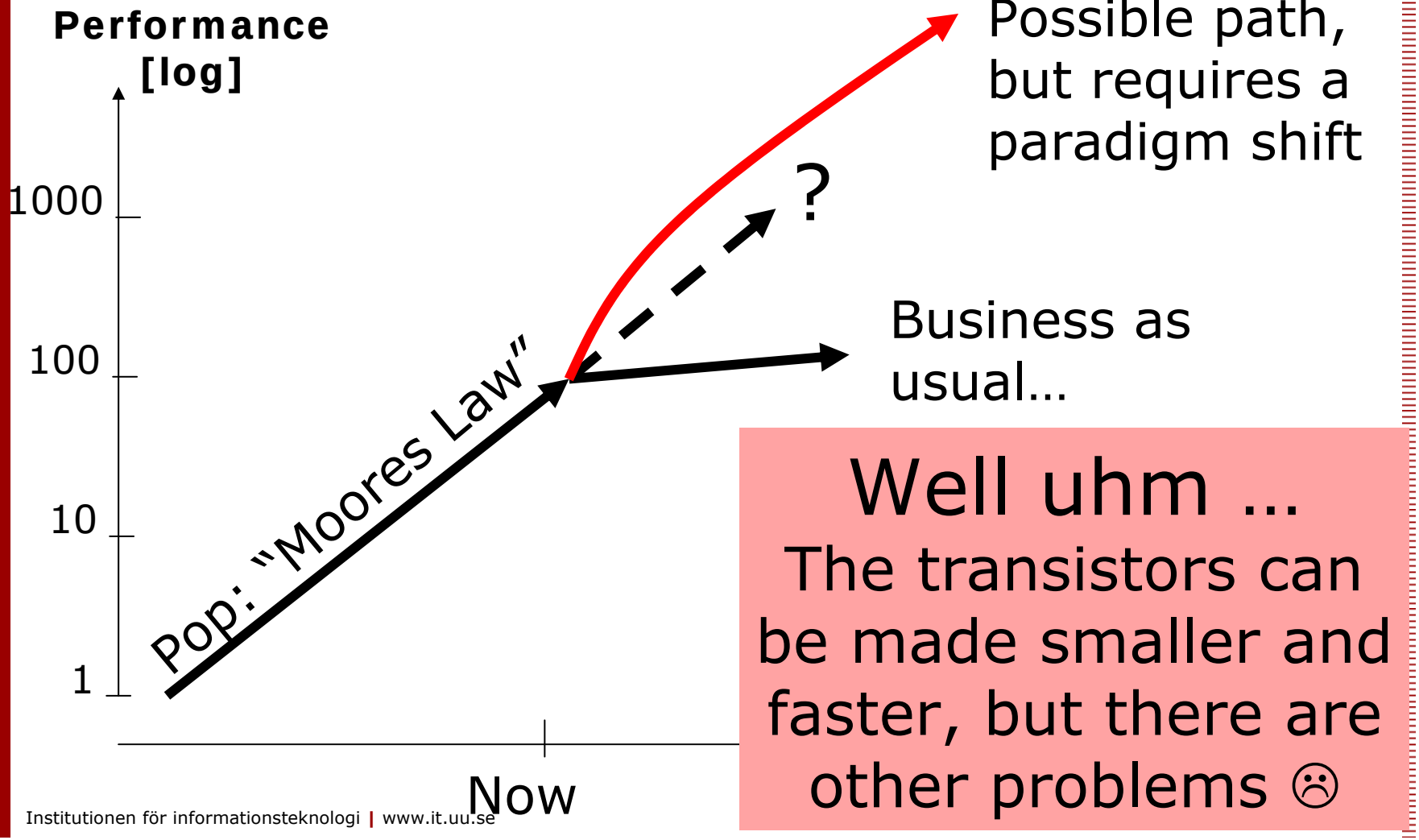


Parallel Threads
a requirement for future CPU chips

Erik Hagersten
Uppsala University
Sweden

Are we hitting the wall now?

Pop: Can the transistors be made even smaller and faster?

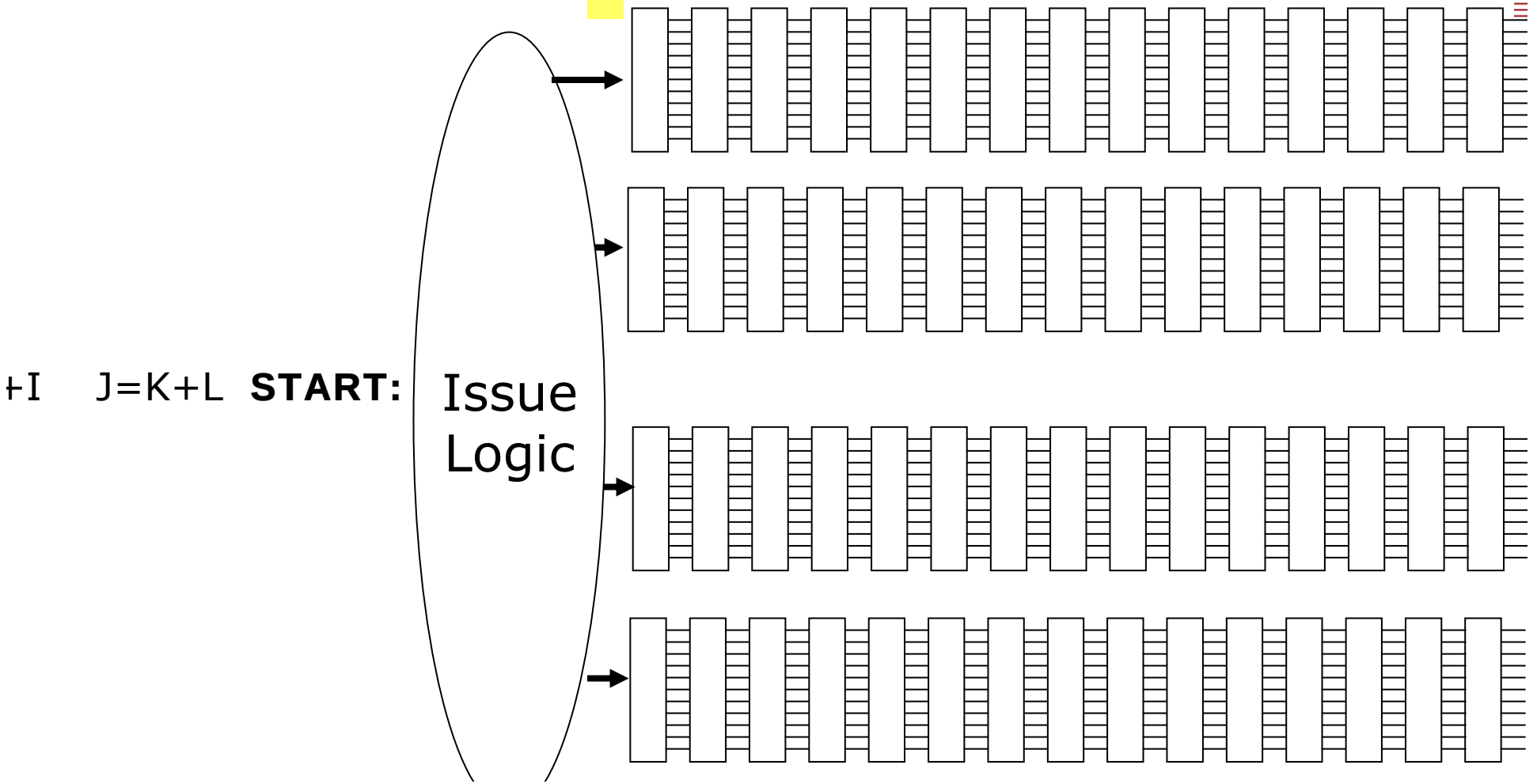


Well uhm …

The transistors can be made smaller and faster, but there are other problems ☹

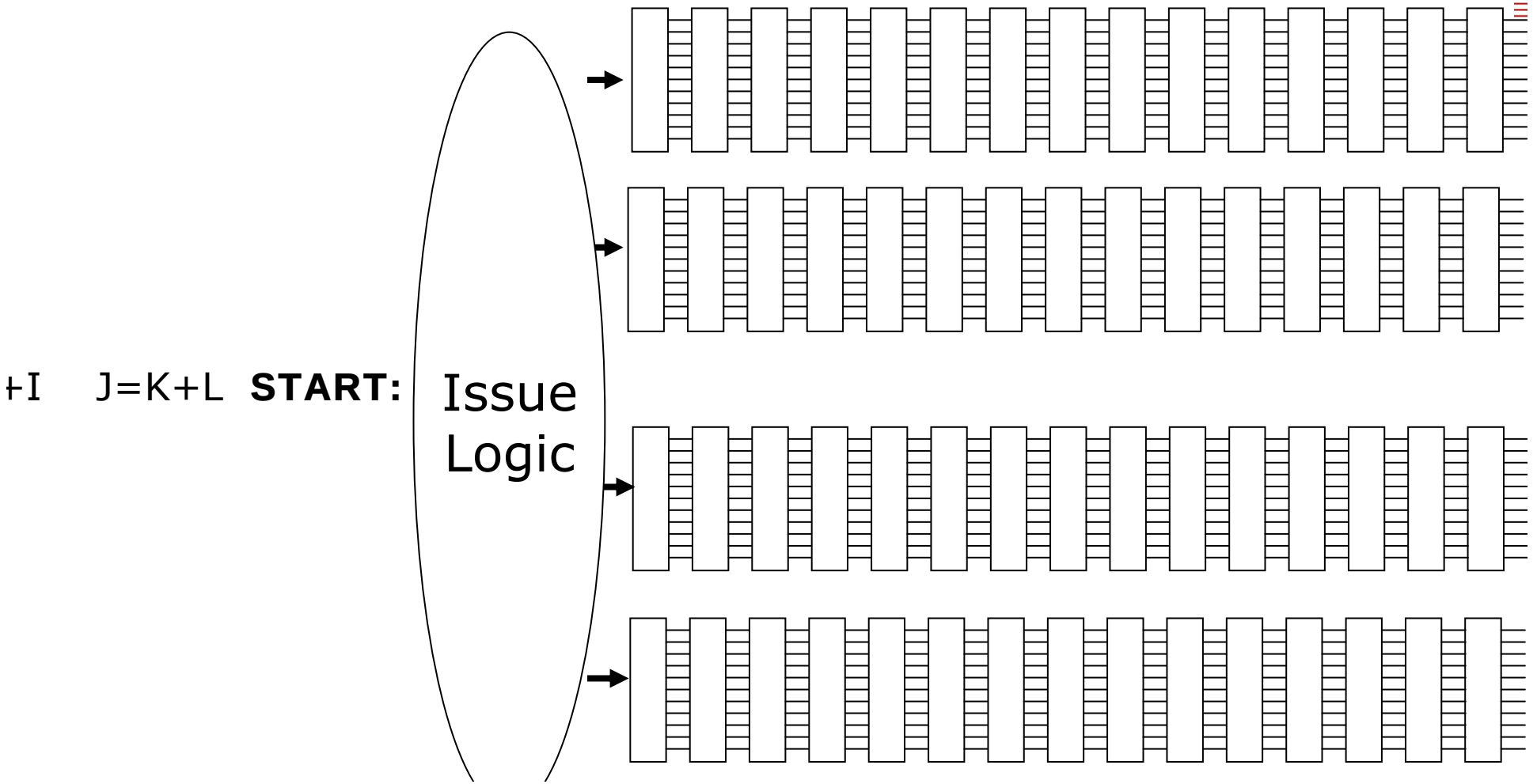
Instruction-Level Parallelism (ILP) in Superscalar Pipelines

```
START:
  J=K+L
  G=H+I
  D=E+F
  A=B+C
  MEM[X]=MEM[X]+14
  X=X+1
  IF X < 1000 GOTO START:
```



Instruction-Level Parallelism (ILP) in Superscalar Pipelines

```
START:
  J=K+L
  G=H+I
  D=E+F
  A= B+C
  MEM[X]=MEM[X]+14
  X=X+1
  IF X < 1000 GOTO START:
```



UPPSALA UNIVERSITET

Uppsala University

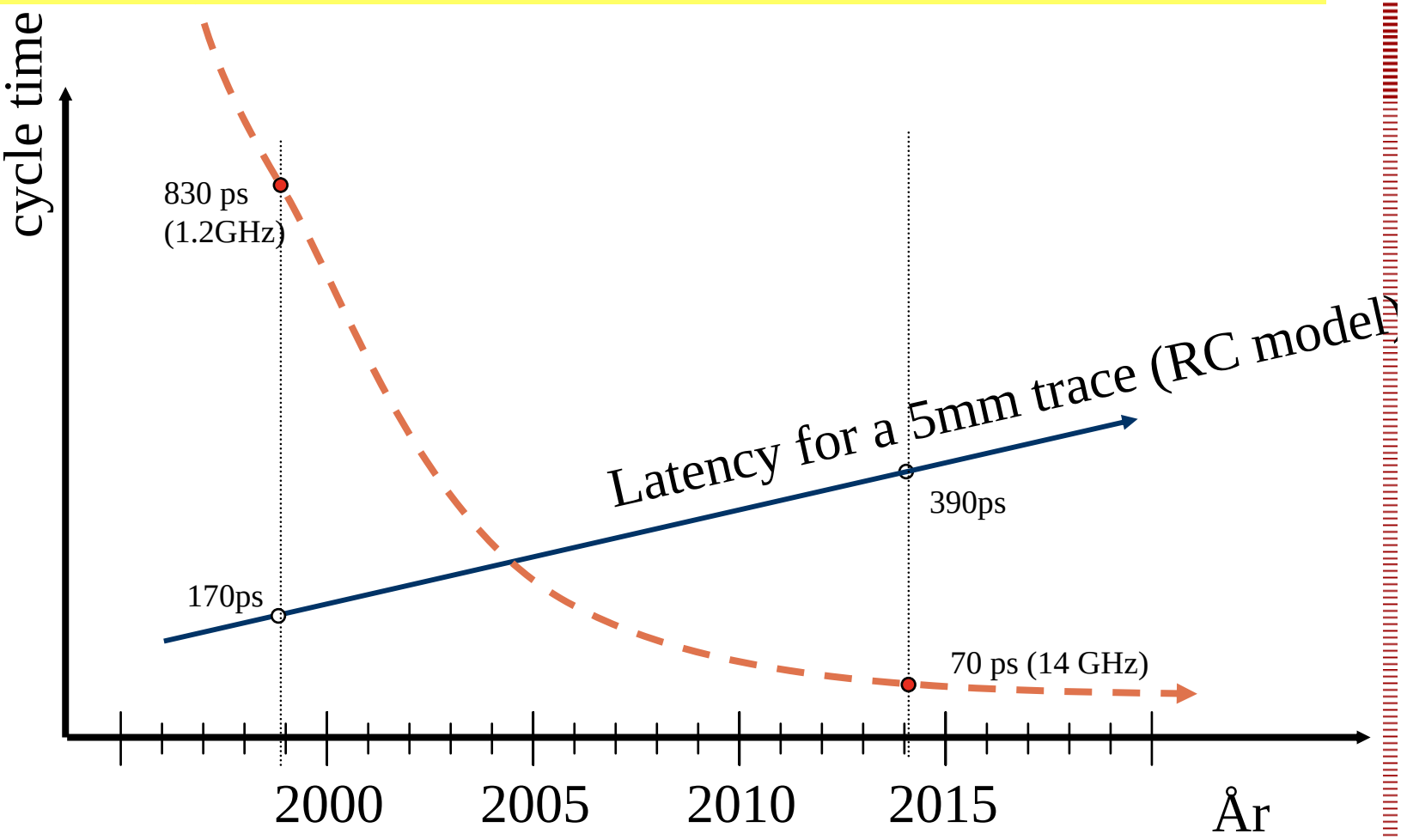
Microprocessors today: What it takes to run one progra

Exploring ILP (instruction-level

- Faster clocks ➔ Deep pip
- Superscalar Pipelines
- Branch Prediction
- Out-of-Order Ex
- Trace Cache
- Speculatio
- Predica
- Adv ess Table
- R tack
-

Bad News #1:
We have already explored most ILP
(instruction-level parallelism)

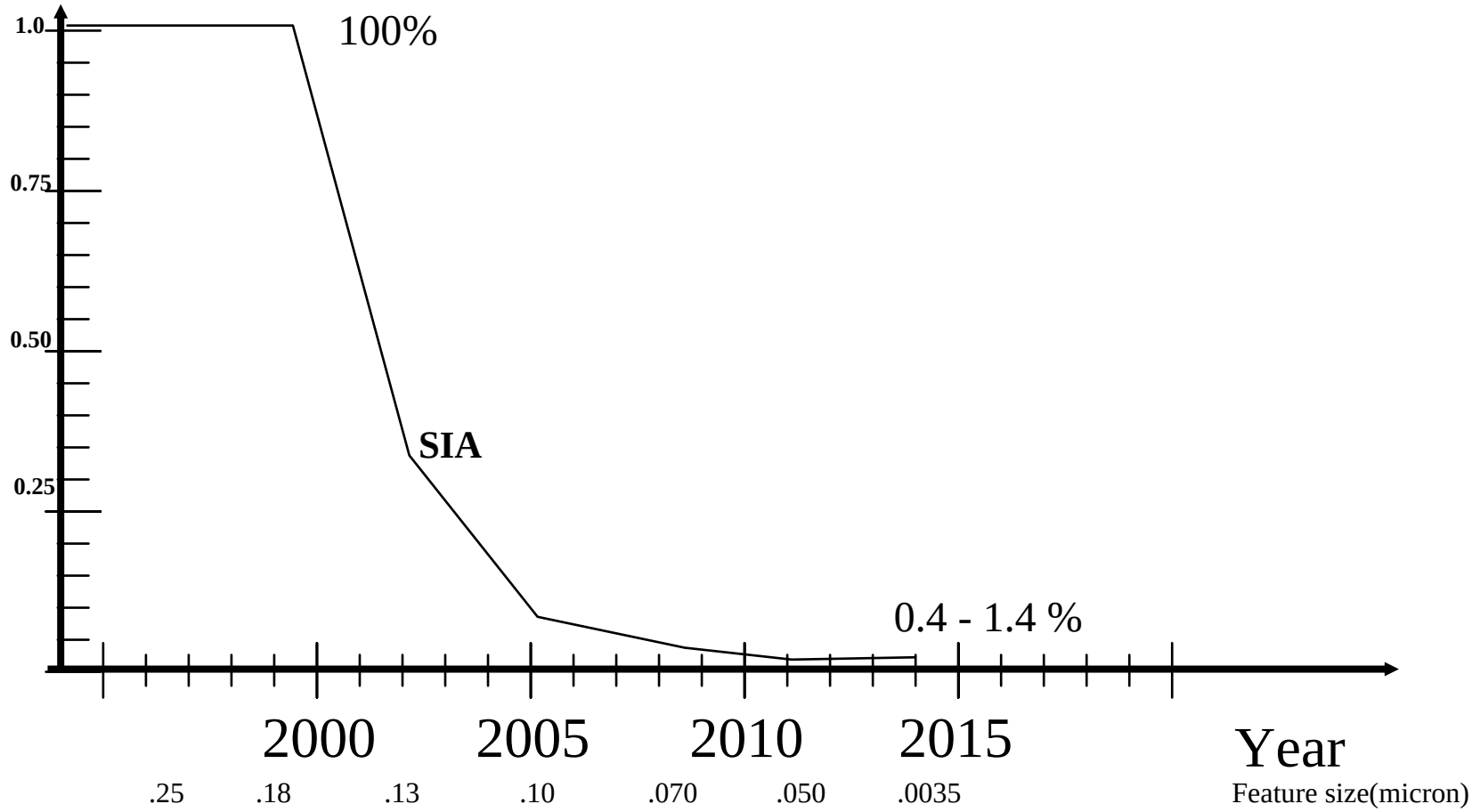
Bad News #2
Looong wire delay →slow CPUs



Quantitative data and trends according to V. Agarwal et al., ISCA 2000
***Based on SIA* (Semiconductor Industry Association) prediction, 1999**

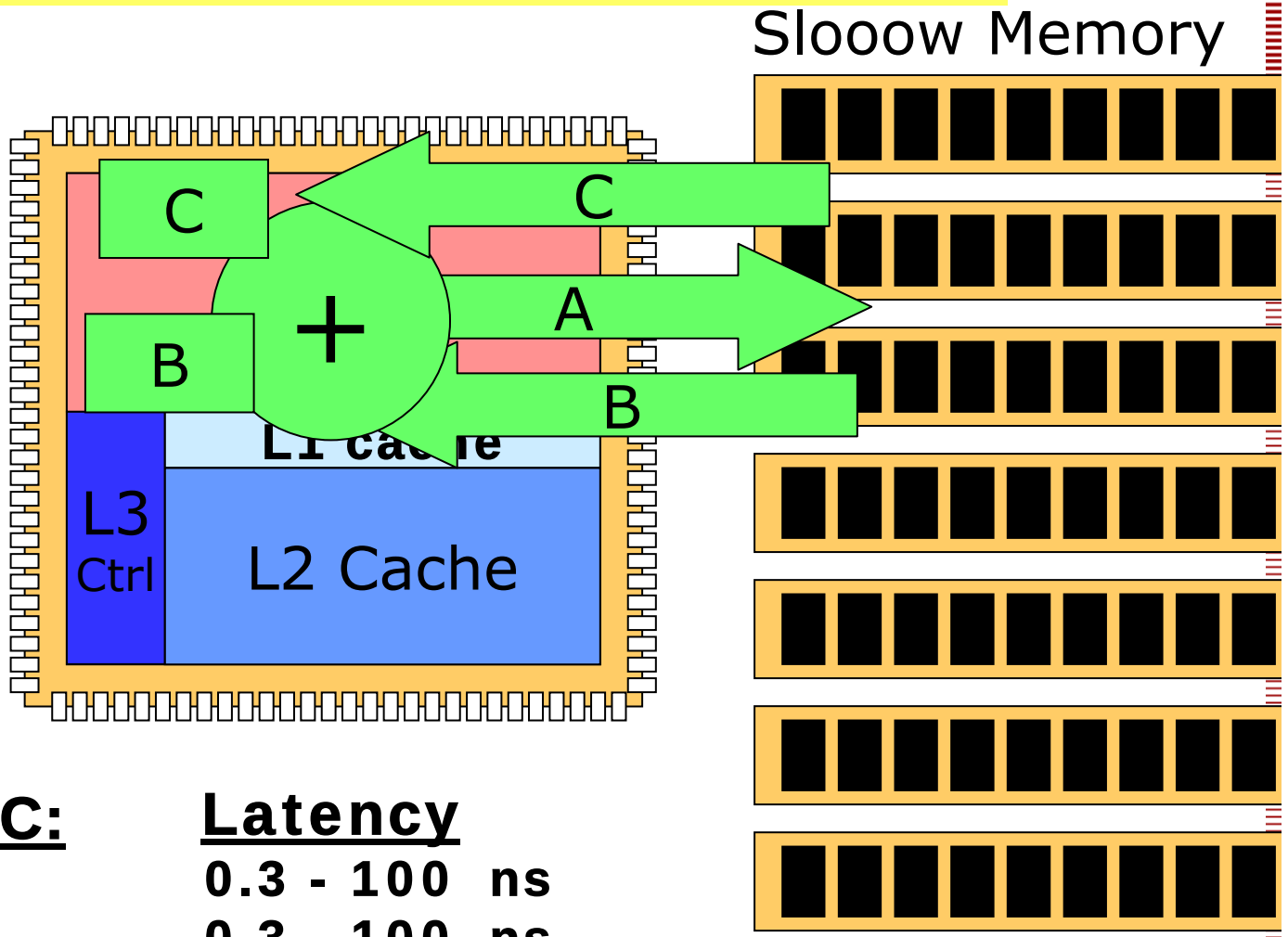
Bad News #2
Looong wire delay →slow CPUs

Span -- Fraction of chip reachable in one cycle



Bad News #3:

Memory latency / bandwidth is the bottleneck…



A = B + C:	Latency
Read B	0.3 - 100 ns
Read C	0.3 - 100 ns
Add B & C	0.3 ns
WriteA	0.3 - 100 ns

UPPSALA UNIVERSITET

Uppsala University

Bad News #4:
Power is the limit

- Power consumption is the bottleneck
  - Cooling servers is hard
  - Battery lifetime for mobile computers
  - Energy is money

- Dynamic effect is proportional to
  - ~ Frequency
  - ~ Voltage^2

UPPSALA UNIVERSITET

Uppsala University

Now What?

#1: Running out of ILP

#2: Wire delay is starting to hurt

#3: Memory is the bottleneck

#4: Power is the limit

UPPSALA UNIVERSITET

Uppsala University

Solving all the problems: exploring threads parallelism

#1: Running out of ILP

➔ feed one CPU with instr. from many threads

#2: Wire delay is starting to hurt

➔ Multiple small CPUs with private L1$

#3: Memory is the bottleneck

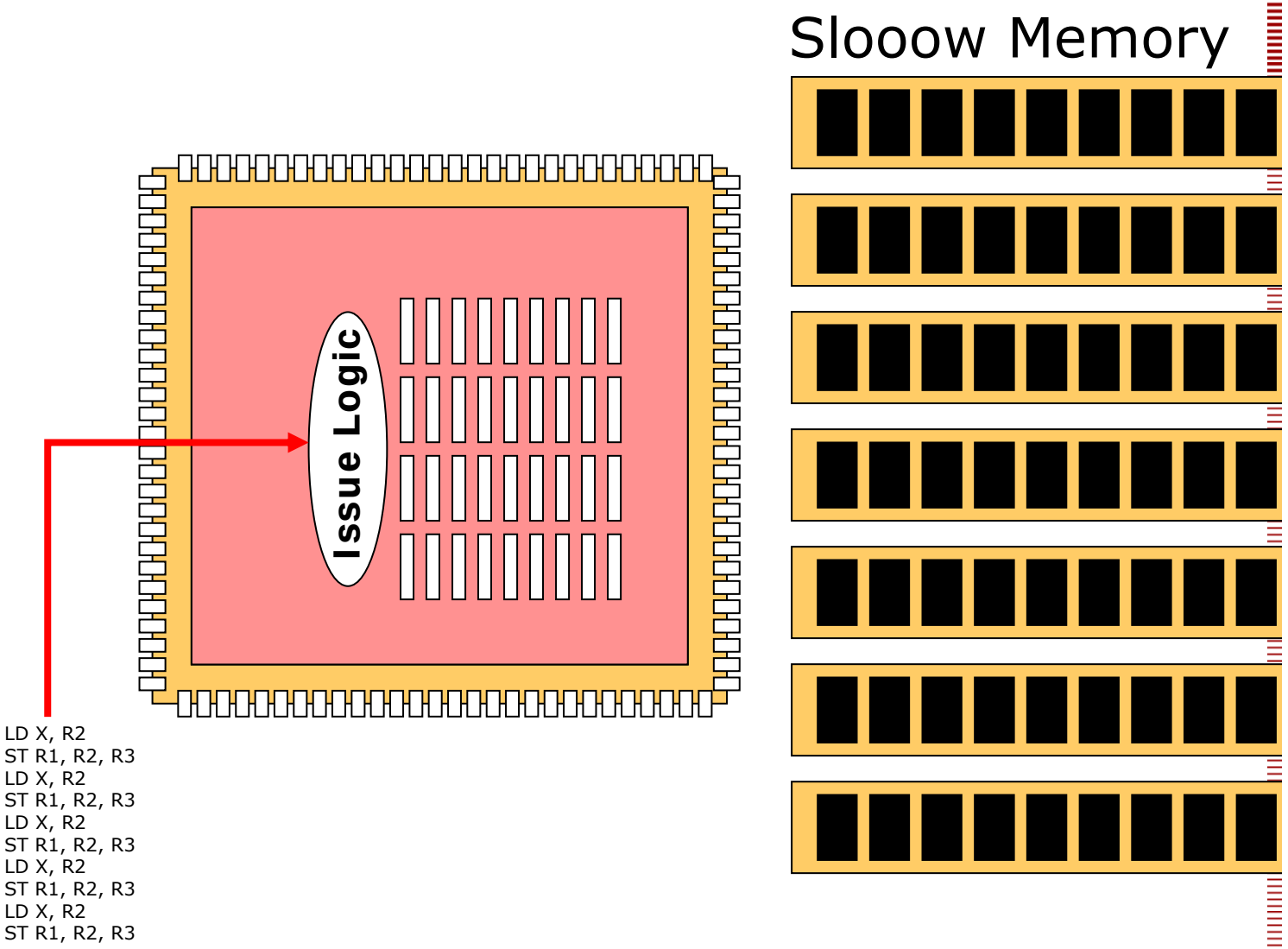
➔ memory accesses from many threads (MLP)

#4: Power is the limit

➔ Lower the frequency ➔ lower voltage

Bad News # 2: Not enough ILP 1(2)

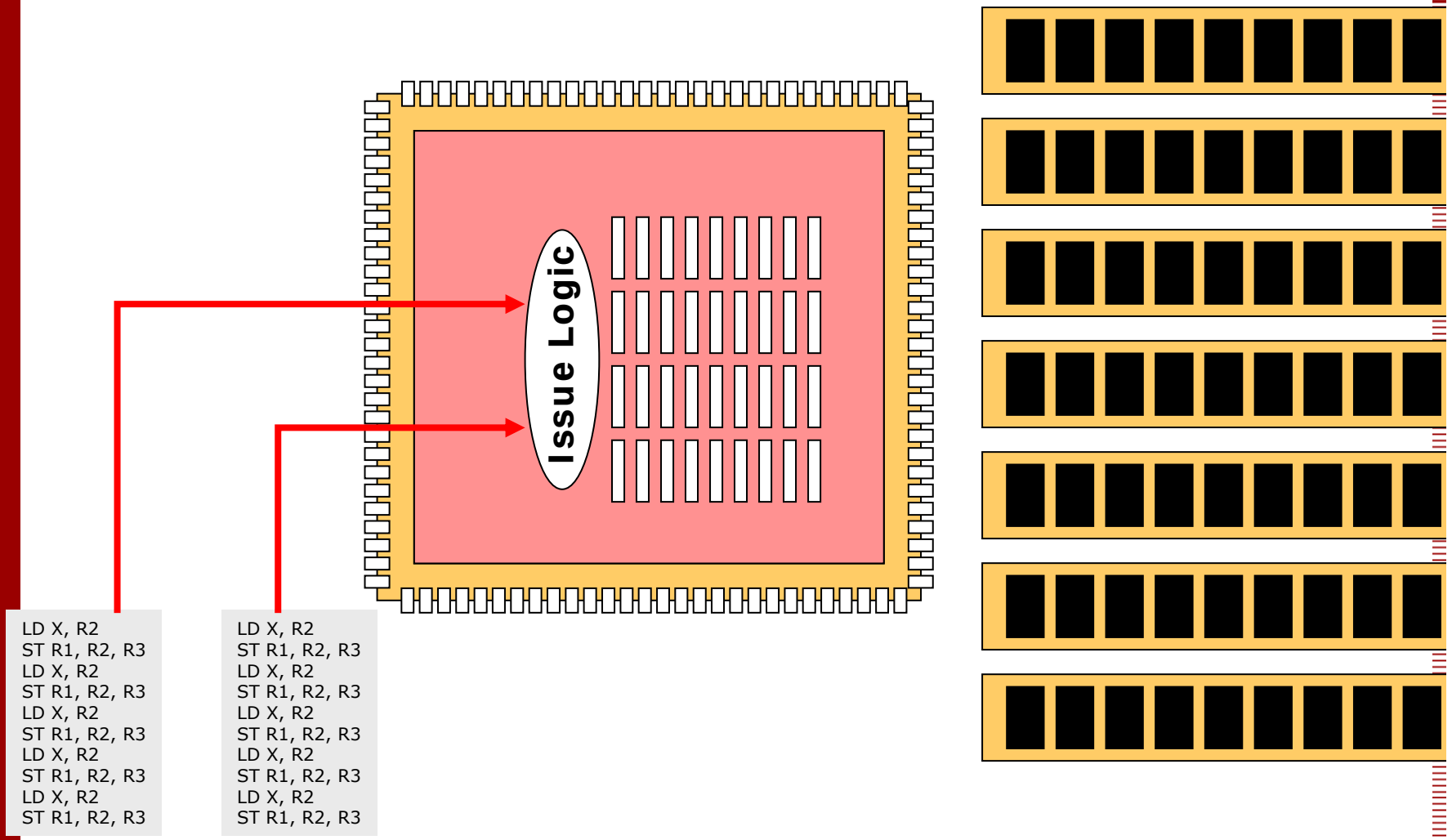
➔ **feed one CPU with instr. from many threads**



UPPSALA UNIVERSITET

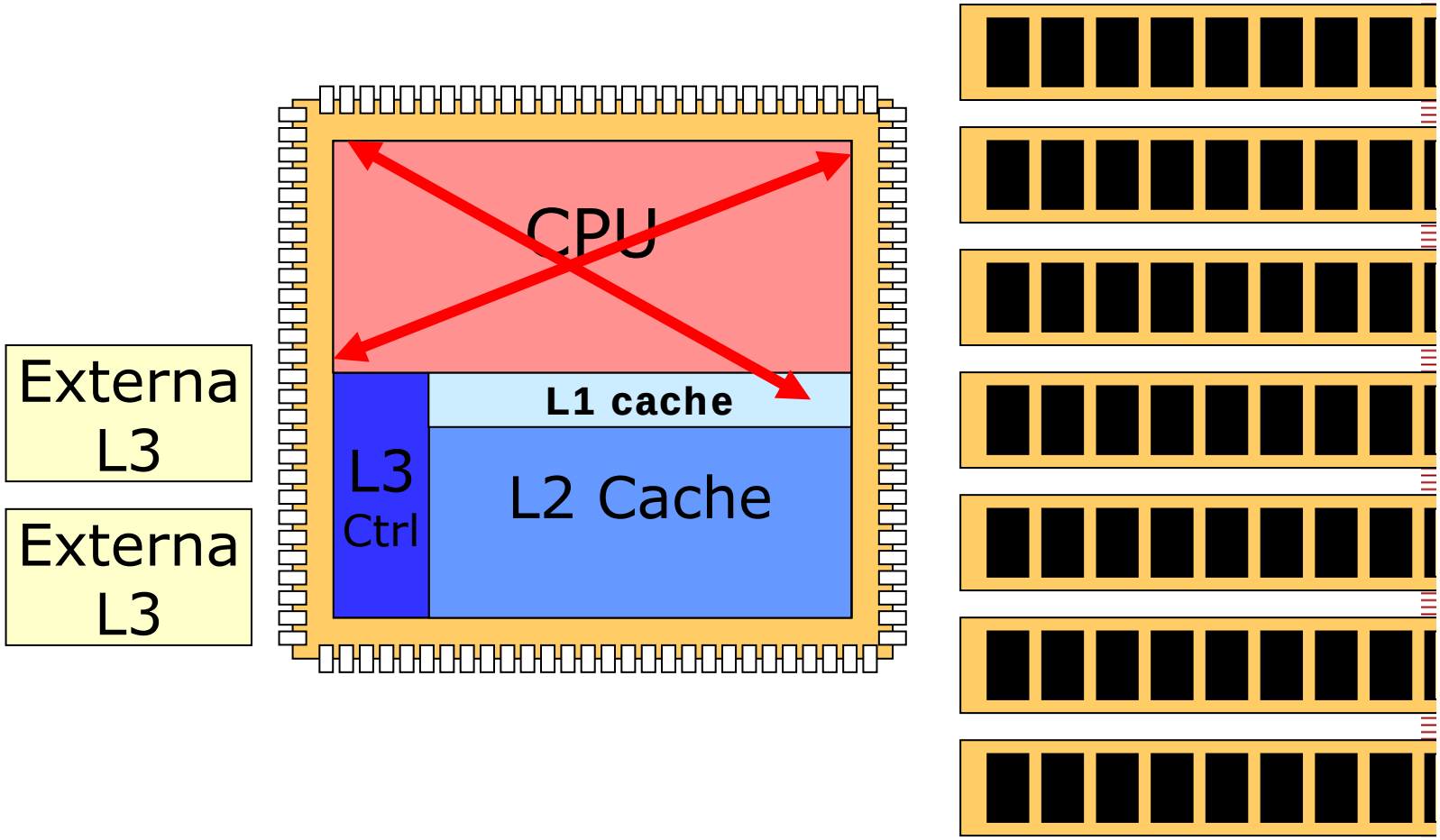
Uppsala University

➔ feed one CPU with instr. from many threads



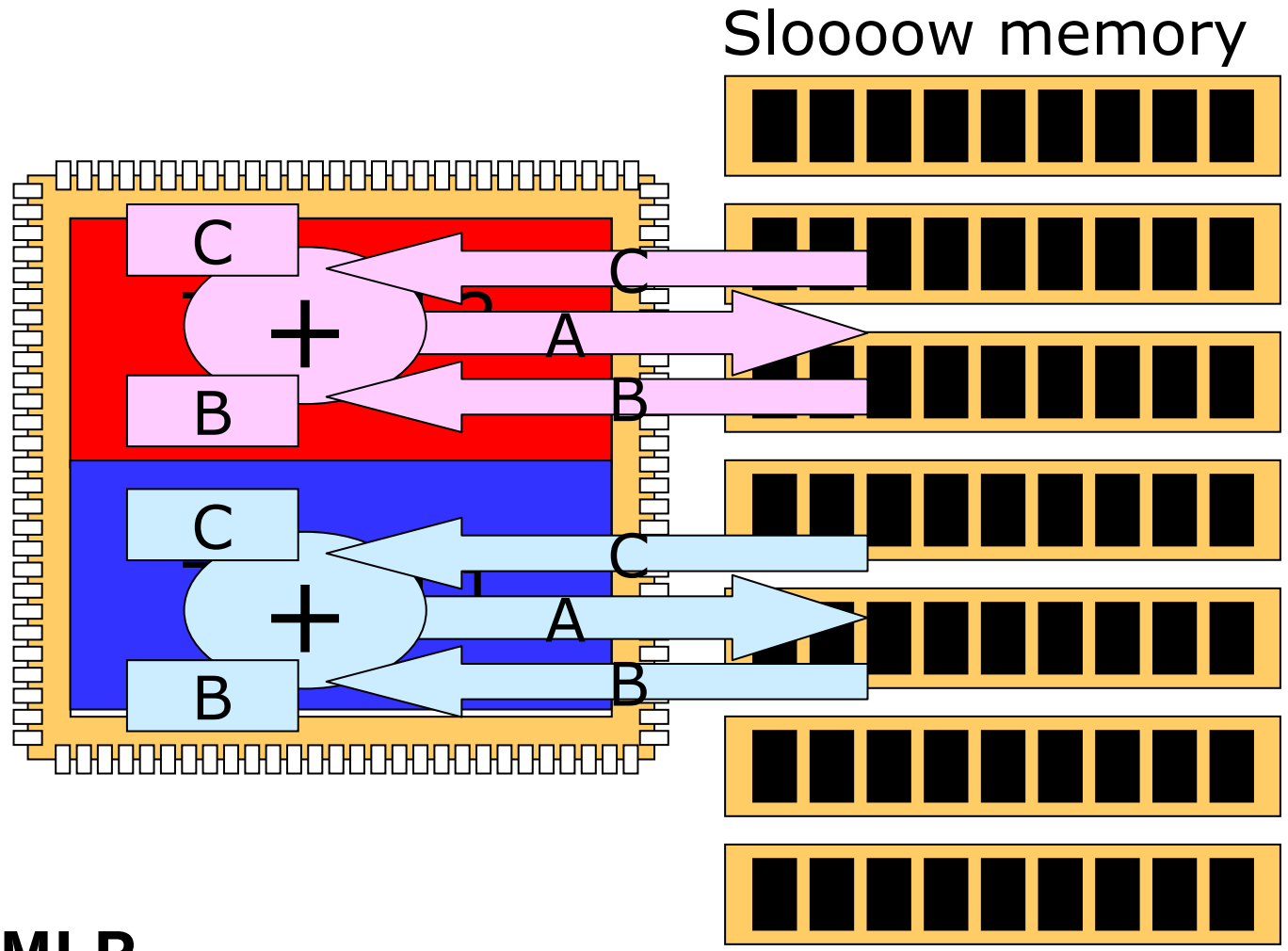
Bad News #2: wire delay

➔ Multiple small CPUs with private L1$



latency / bandwidth

➔ memory accesses from many threads (MLP)



TLP ➔ MLP

Thread-Level Parallellism ➔ Memory-Level Parallelism (MLP)

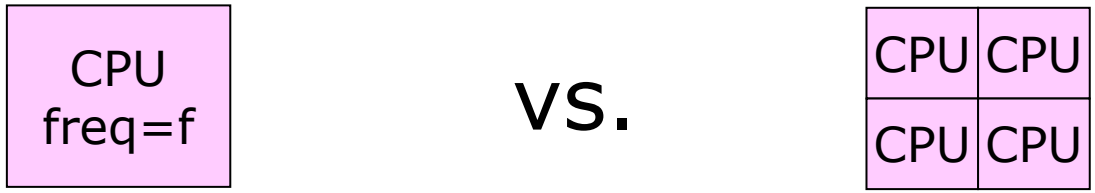
consumption

➔ Lower the frequency ➔ lower voltage

$$P_{dyn} = C * f * V^2 \approx \text{area} * \text{freq} * \text{voltage}^2$$

$P_{dyn}(C, f, V) = CfV^2$

$P_{dyn}(2C, f/2, <V) < CfV^2$



$P_{dyn}(C, f, V) = CfV^2$

$P_{dyn}(C, f/2, <V) < ½\ CfV^2$

Throughput < 2x

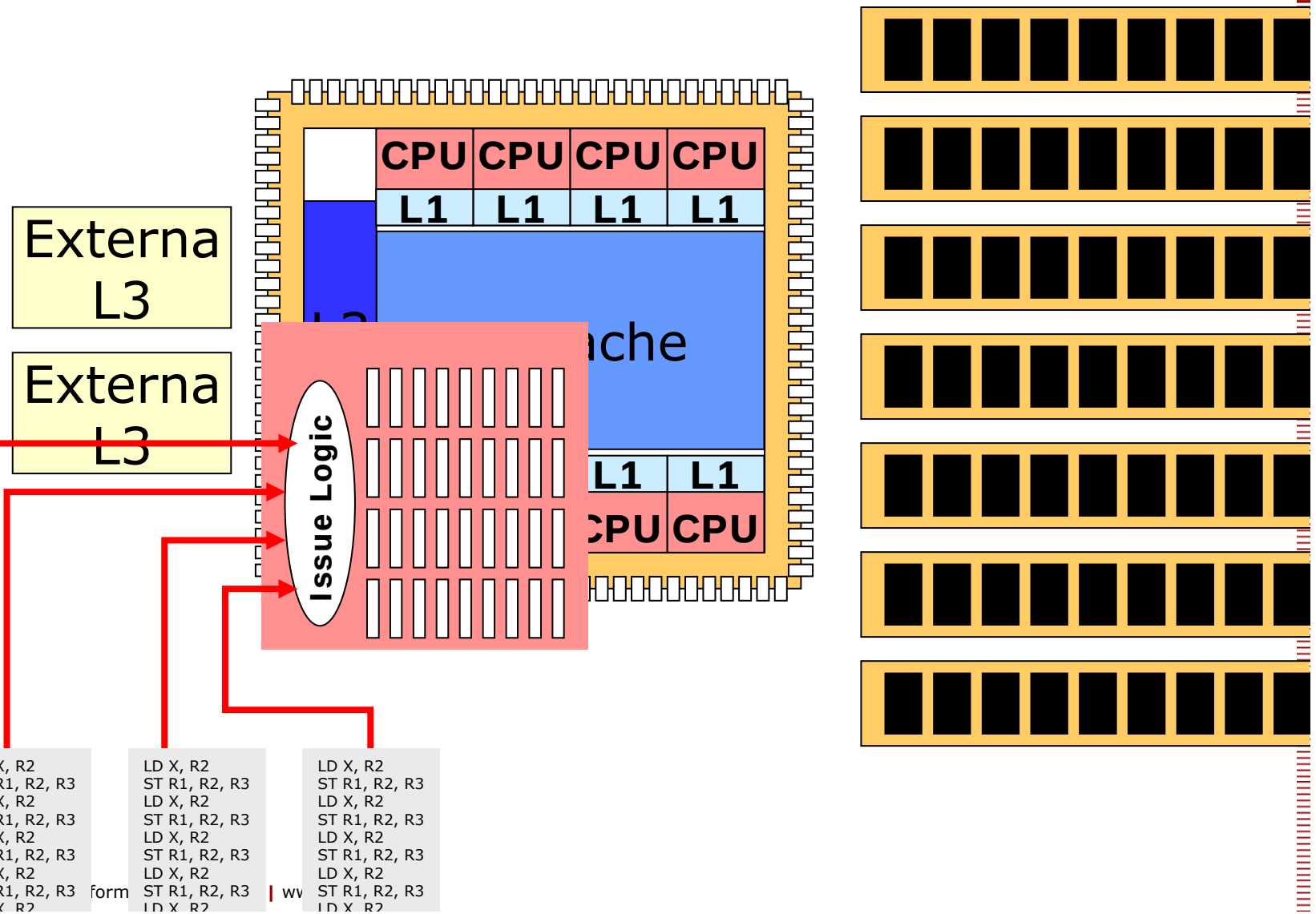
UPPSALA UNIVERSITET

Uppsala University

Around the corner…

Chip Multiprocessors (CMP)

Simultaneous Multithreading (SMT)

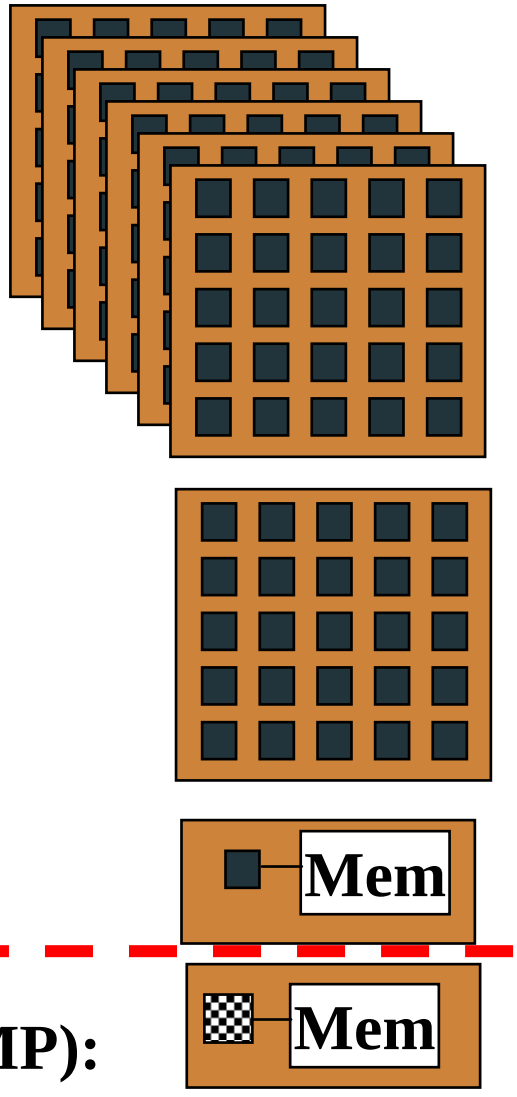


Darling, I shrunk the computer

Mainframes

Super Minis:

Microprocessor:



Paradigm Shift

Chip Multiprocessor (CMP):
A multiprocessor on a chip!

UPPSALA UNIVERSITET

Uppsala University

Now, everyone is doing it!

"Intel have 10 projects in the works that contain four or more computing cores per chip"
[Paul Otellini, Intel Chief Excecutive at IDF fall 2005]

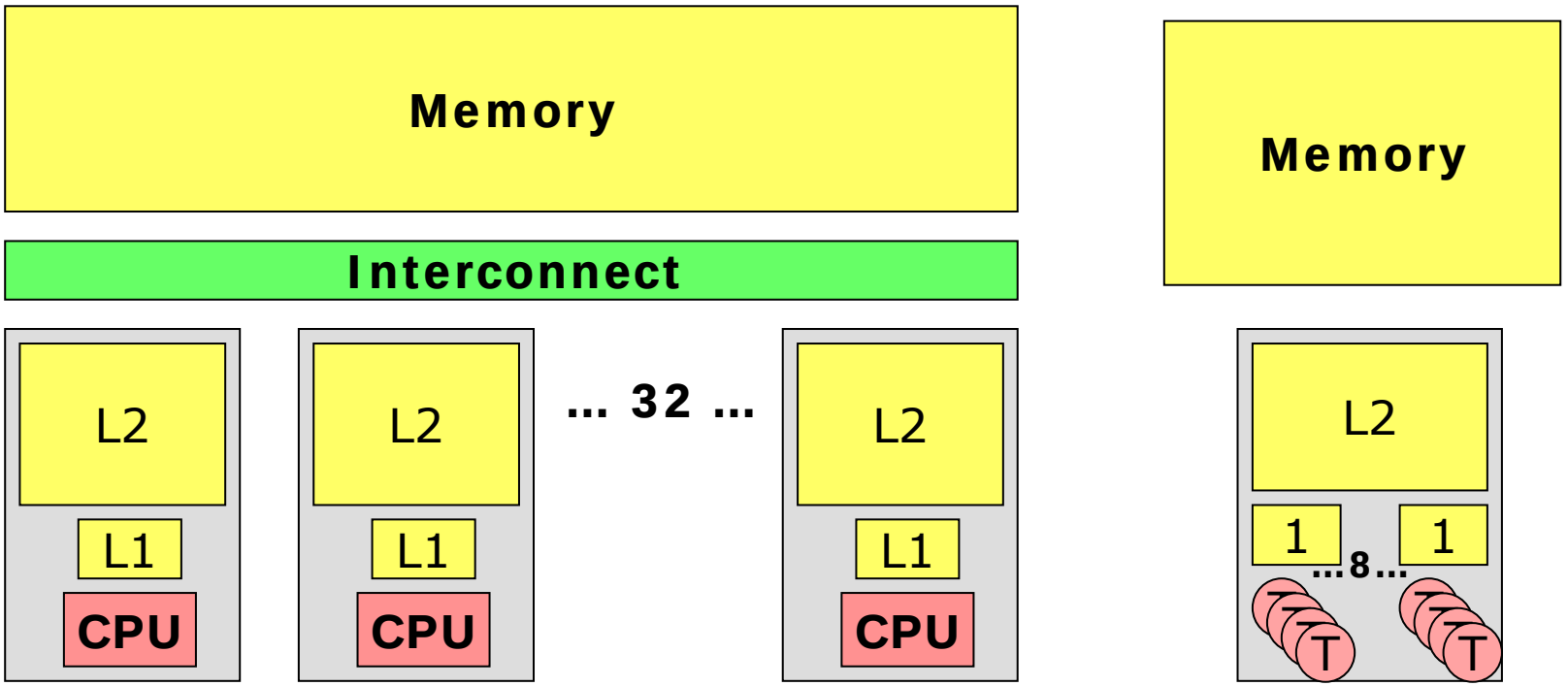
"Today, processors with multiple CPUs and a large cache on a single chip are becoming common

Attempts to tease the parallelism out of a sequential program automatically haven't worked out very well

We need better education, better languages, and better tools, since building concurrent programs is hard"
[Andrew Herbert, Director of Microsoft Cambridge Research Lab, May 2005]

How to get parallelism?

Looks and Smells Like an SMP?



Well, how about:

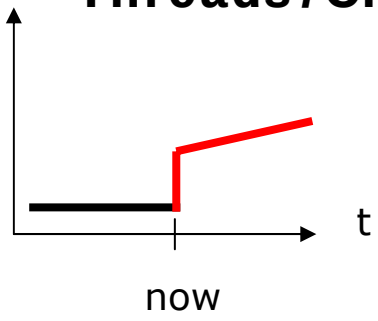
- Cost of parallelism?
- Cache capacity per thread?
- Memory bandwidth per thread?
- Cost of thread communication? …

UPPSALA UNIVERSITET

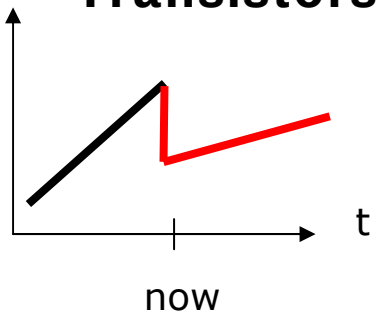
Uppsala University

Trends (my guess!)

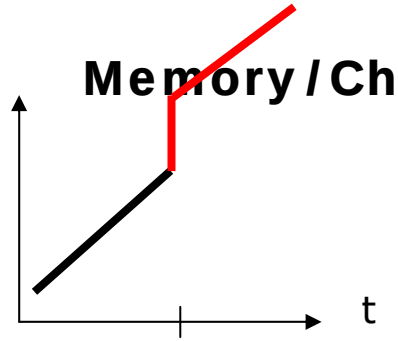
Threads / Chip



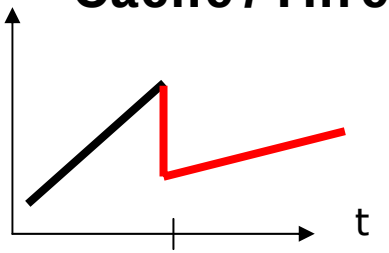
Transistors / Thread



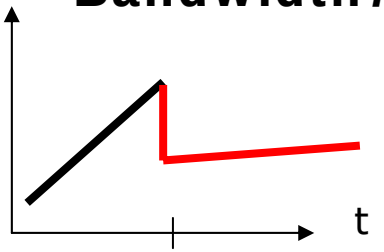
Memory / Chip



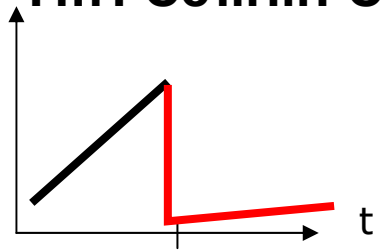
Cache / Thread



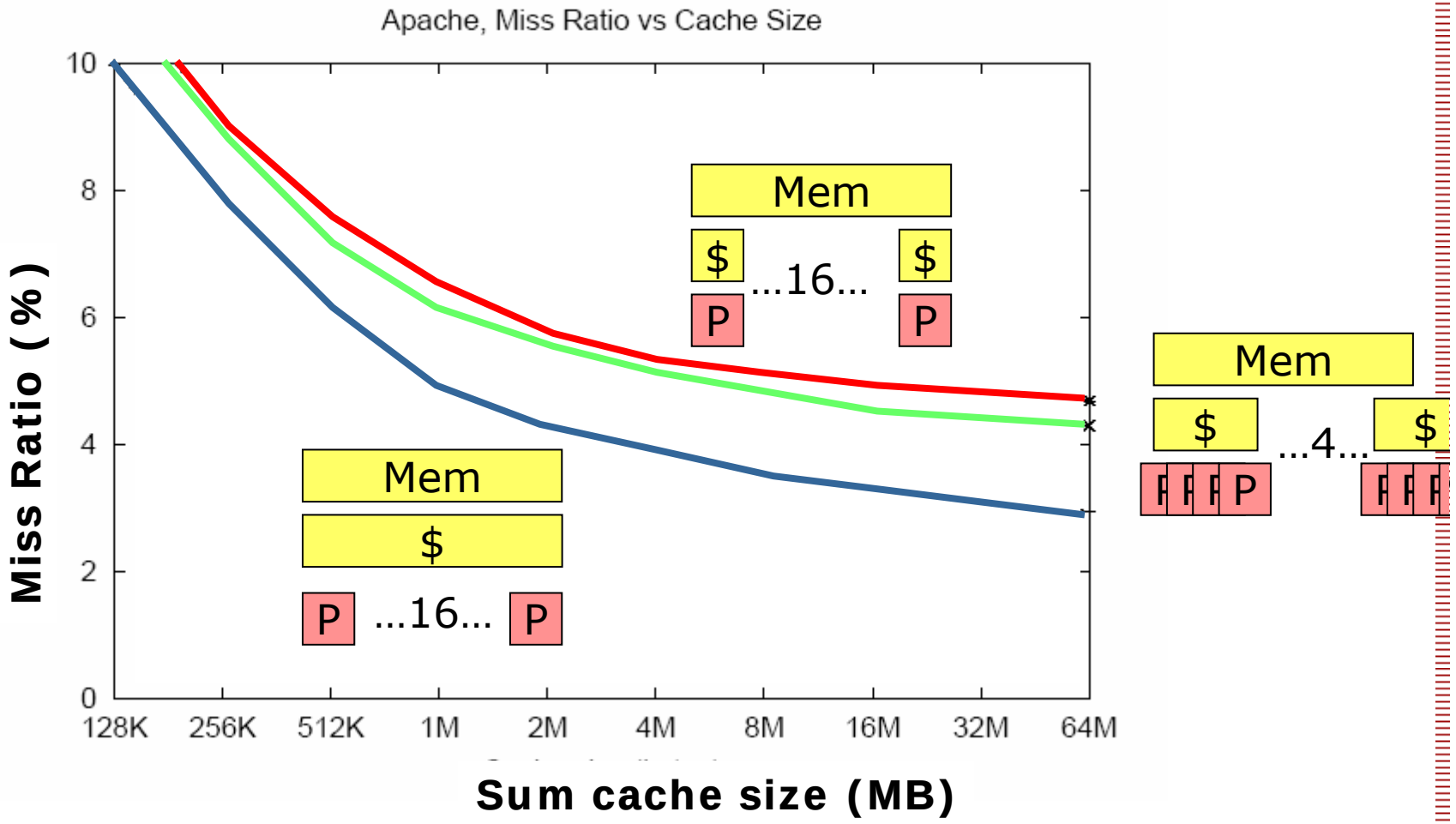
Bandwidth / Thread



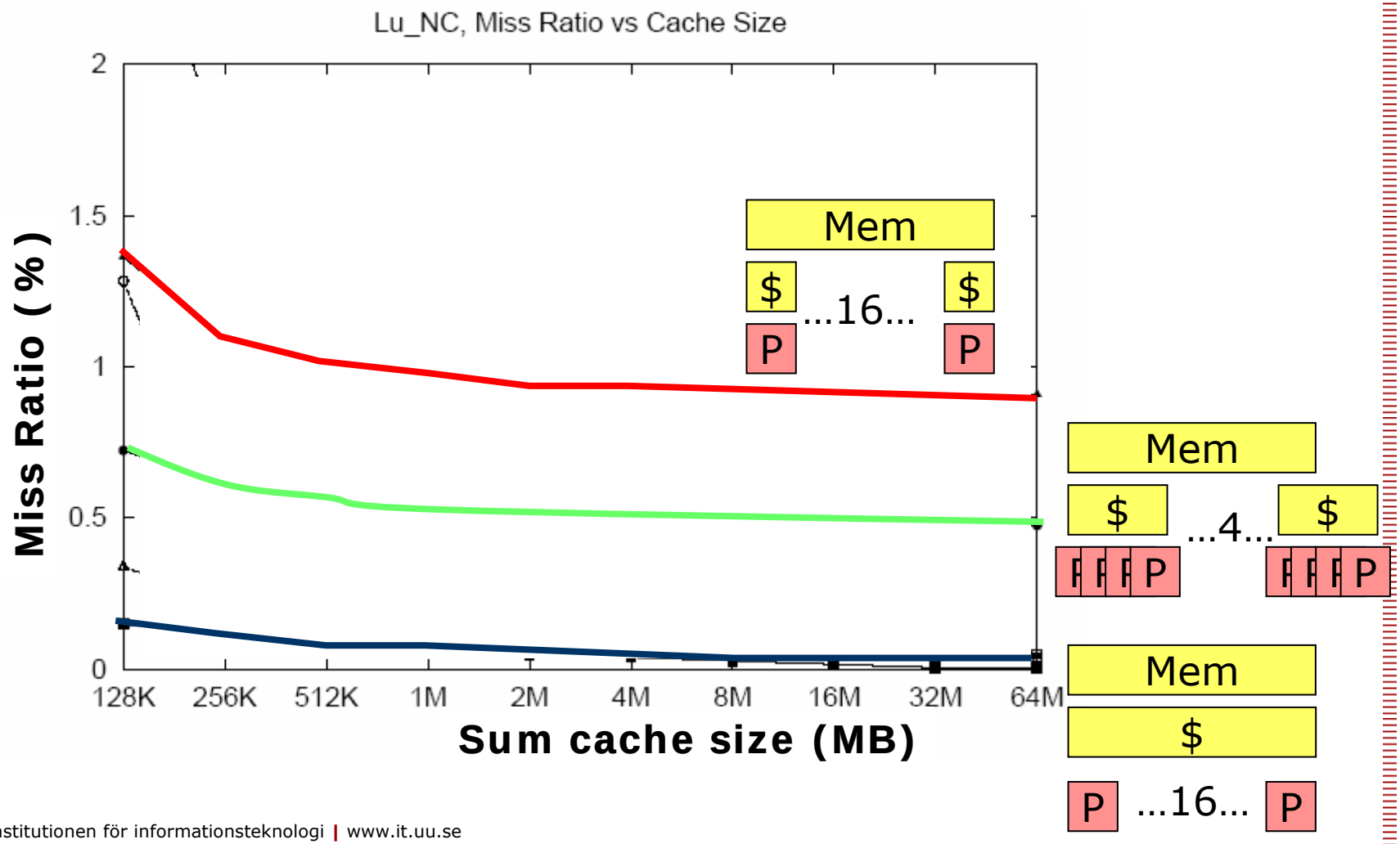
Thr. Comm. Cost (temporal)



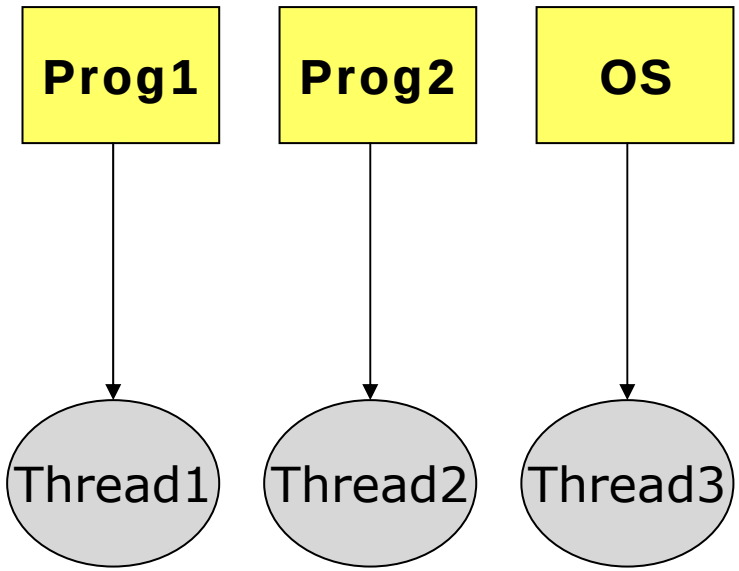
Apache, 16 threads



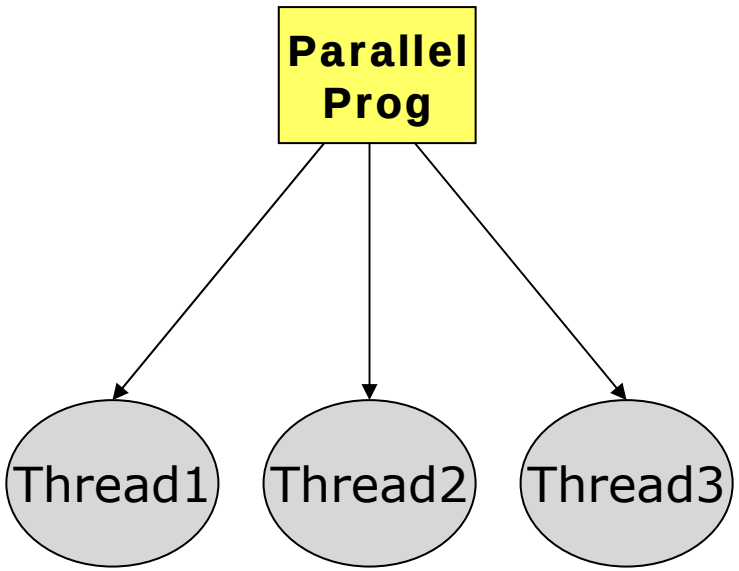
LU_NC, 16 threads



What thread parallelism?



Capacity Computing



Capability Computing

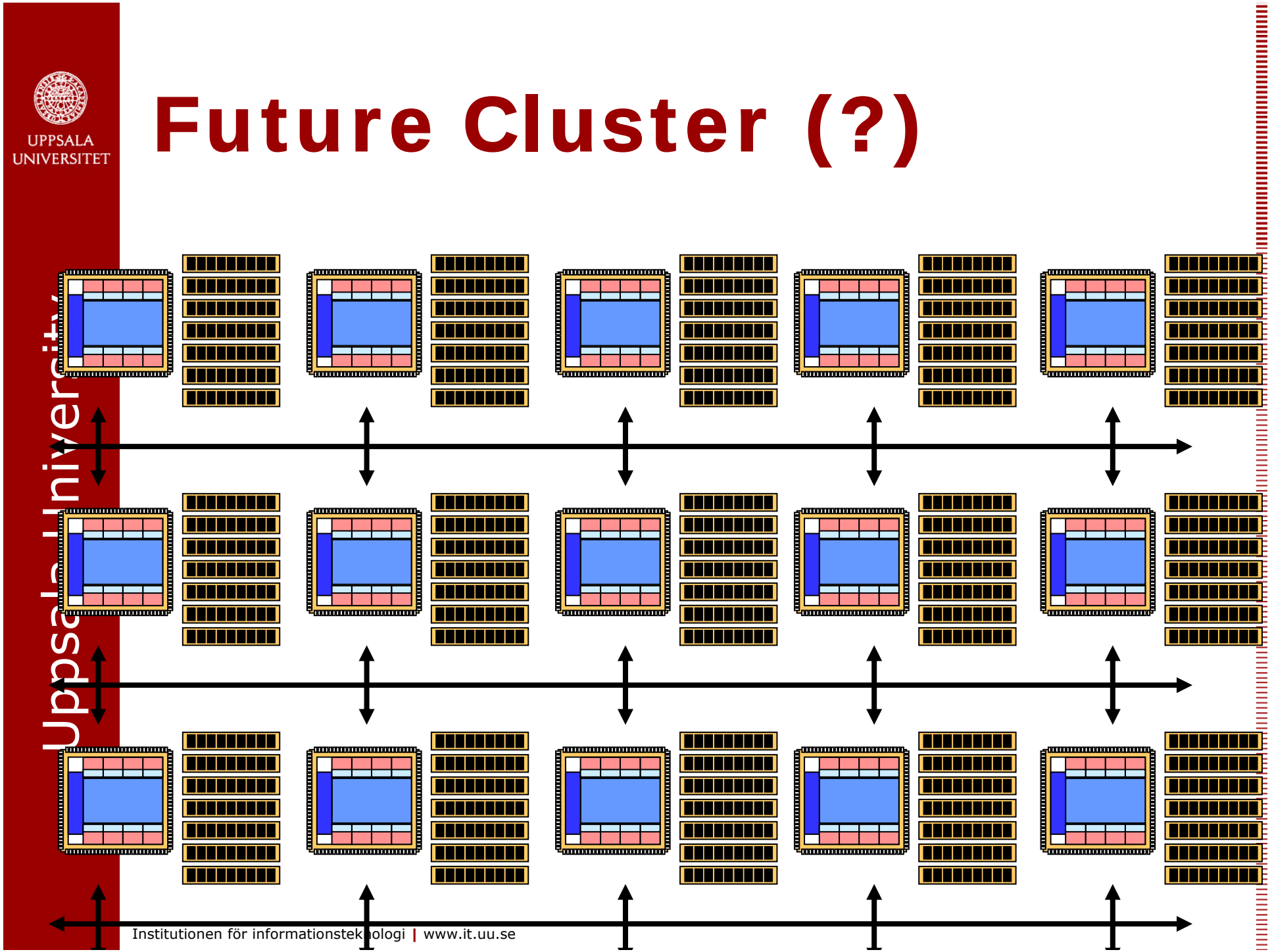
UPPSALA UNIVERSITET

Uppsala University

Questions for the Future

- What applications?
- How to get parallelism and data locality?
- Will funky languages see a renascence?
- Will automatic parallelizing return?
- Are we buying:
  - compute power,
  - memory capacity, or
  - memory bandwidth?
- Will the CPU market diverge into desktop/capability CPUs again?

- A non-question: will it happen?


UPPSALA
UNIVERSITET
Future Cluster (?)
Uppsala University

UPPSALA UNIVERSITET

Uppsala University

CMP Clusters

- Great potential
  - More then "Moore's"
  - Great performance/watt
- Node model: Buy one thread, get 31 for free!
- Less cache per thread
- Memory bandwidth is THE bottleneck
  - ➔ Does not speak in favor of capacity computing

➔ Capability computing within each node

➔ Capability/Capacity within the cluster

➔ Mediocre node speedup is OK (it's for free)

UPPSALA UNIVERSITET

Uppsala University

Summing up: Missing the wall

- Thread-level parallelism in all processors
- New walls: Bandwidth & parallelism
- Parallel software is tough stuff!
- Important R&D area (again)
  - Algorithms
  - Parallelization
  - Verification
  - Modeling/Simulation/Tools
  - Keep bandwidth under control
  - Managing locality