Application specific memory network organizations for nanometer technologies

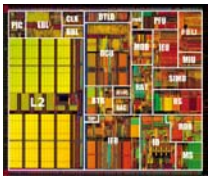
Rudy Lauwereins

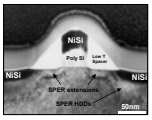
Vice-President IMEC, Belgium
Professor at Katholieke Universiteit Leuven, Belgium

SEEDS FOR TOMORROW'S WORLD

Design in 45nm and beyond: interconnect case study

End-user value



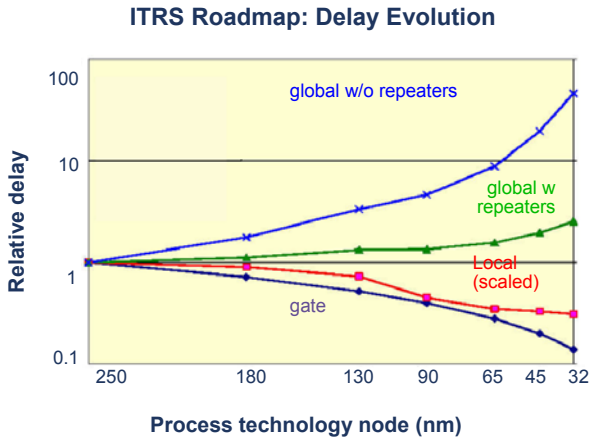


Scaling threats:

- Interconnect
- Variability
- Leakage

Inefficient interconnect scaling endangers good system design

Wires no longer follow the propagation delay of gates: severe delay in long lines that do not scale well (i.e., busses but also wires inside big blocks)



Inefficient interconnect scaling endangers good system design

Technology approaches:

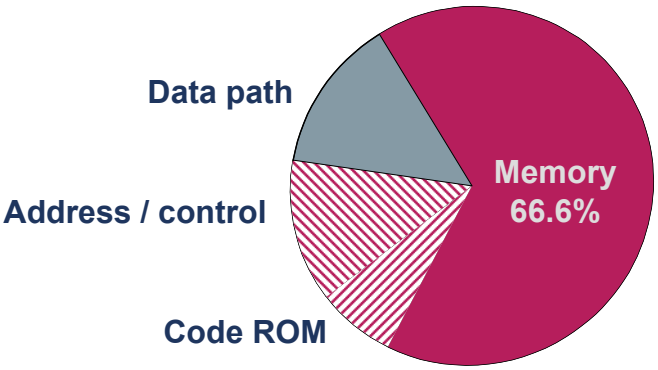
- Low-k and copper research: limited solutions
- But: impact of long lines will keep increasing with technology node: SoC wires do not scale in size

Circuit approaches:

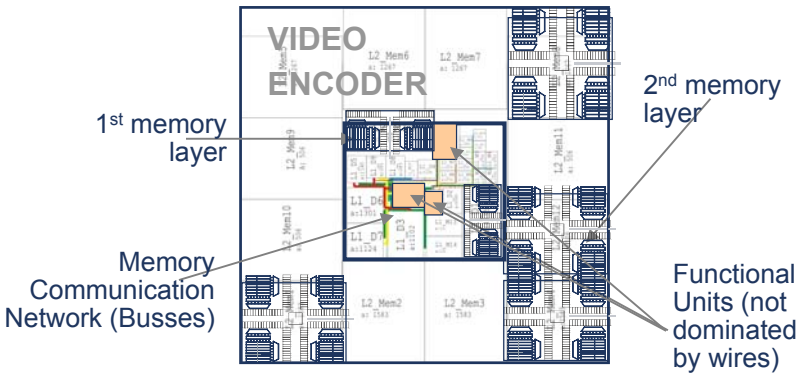
- Add repeaters in long lines to satisfy timing closure
- But: increasing energy overhead in repeaters

Energy efficient system-level solution must solve timing closure problem, or the SoC will not work

Memories dominate power consumption



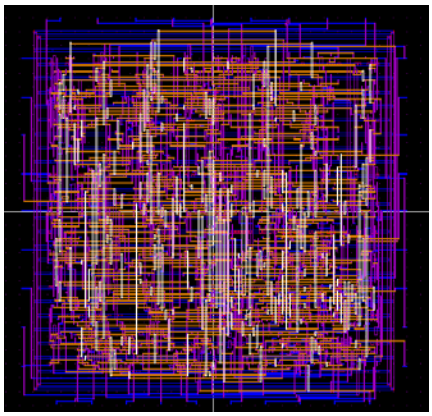
Memory dominated SoCs are full of wires



Interconnect dominates scene but timing closure is feasible:

- Wires inside memories (timing closure solved once per block)
- Wires between memories (only 10-100 blocks for every application, easier to solve)

Why focus?: Standard cell design have irregular floorplans



Wire-length only known after Place&Route: not suited for accurate estimations

Timing closure problem must be solved for each standard cell (1000´s) and for each application!!

State-of-the-art overview

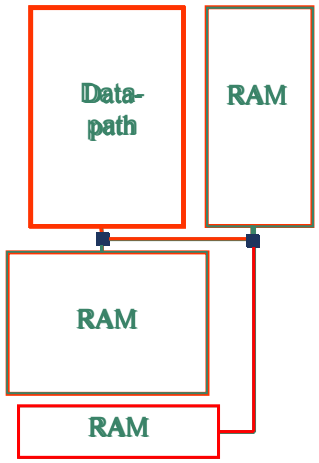
Communication network architectures:

- point to point connections
  - Not popular: not programmable and complex in wiring (congestion)
  - locally optimal for power, hard to solve timing closure
- shared bus
  - Simple in wiring and programmable
  - worst for power, timing closure easier to solve at the expense of power
- segmented bus
  - Evolved from shared buses: [Chen91, Kulick93, Aras92, Aldworth99, Chen99]
  - Slightly more complex than shared bus, but programmable and low power

Macro placement and floor-planning:

- aiming at timing closure, minimizing area and interconnect cost [Sarrafzadeh96, Tang01]
- bus power is improved by placing blocks close to each other thus minimizing wire length of highly active buses [Holt96,Jimenez01], hence aiming at reducing energy consumption
- global system-level approach is needed, combining application activity knowledge with communication network architectures and floor-planning

Bus segmentation: principle



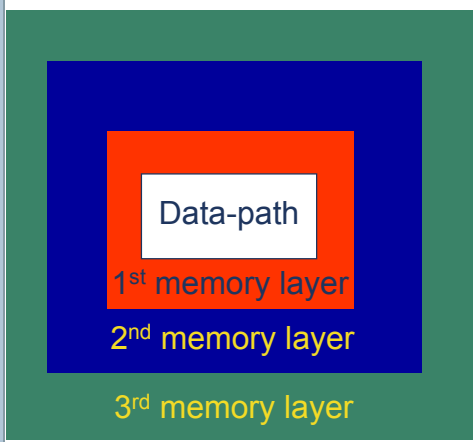
The bus is divided in segments by switches

Only the necessary segments are activated each time

Bus arbitration largely decided at design time based on application requirements

Activity-aware memory floorplan template

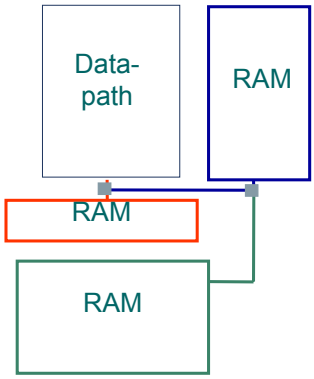
Heavily active memories should be placed closest to the data-path (shorter connections)



Activity
high
medium
low

Heavily active memories are small (due to optimizations), so they are easier to place close to datapath

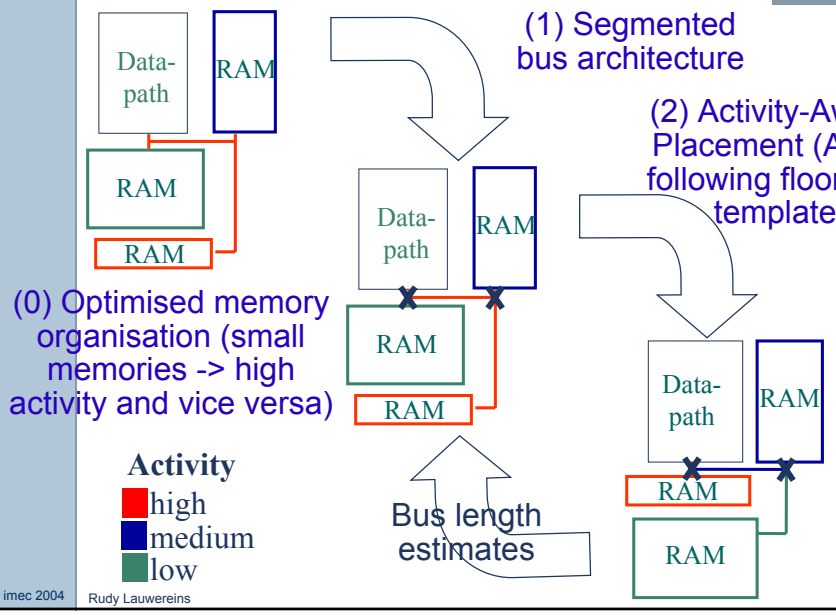
Activity of blocks can be exploited to reduce energy consumption



Memories with higher activity (also smaller if mapping well done) should be placed close to Functional-Units

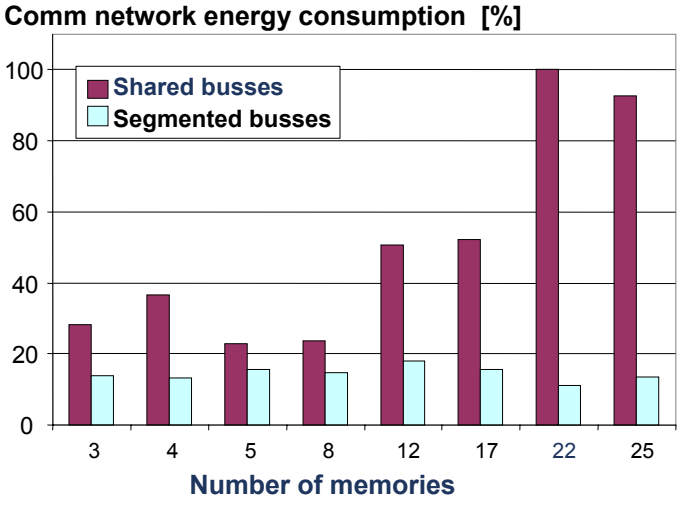
Segments with higher activity get shorter lengths (less load), leading to a globally better solution

A global approach for energy efficient bus design



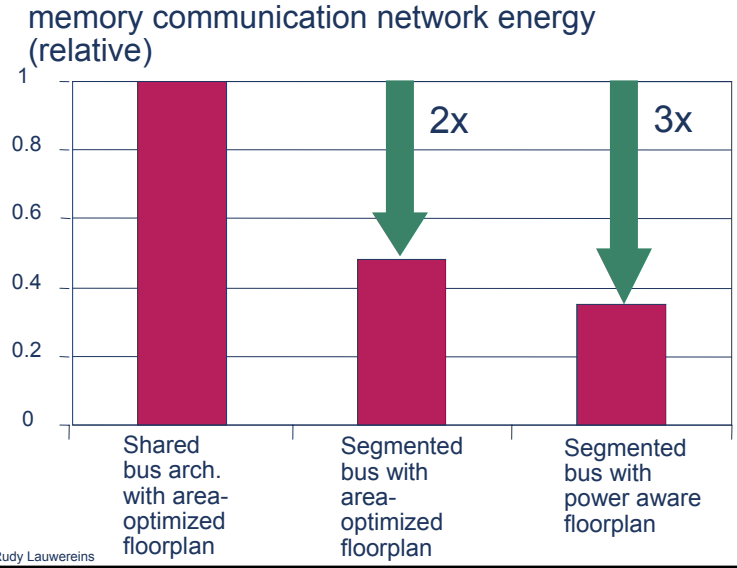
Segmented busses scale much better than shared busses

Hierarchical video encoder



Rudy Lauwereins
13

Impact of floorplanning and bus segmentation techniques on DAB

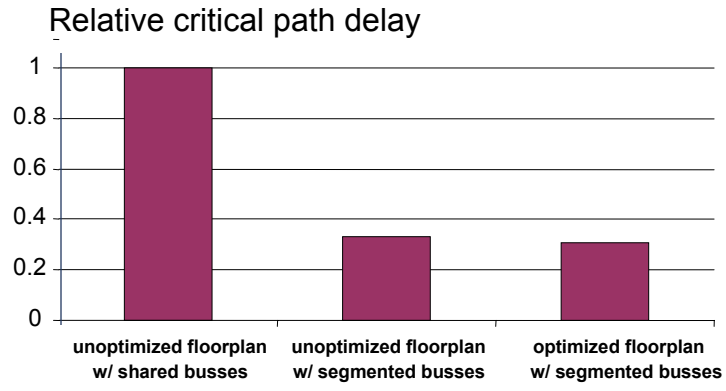


Rudy Lauwereins
14

Impact on area is negligible, delay is improved

- Area overhead is less than 5% compared to reference design for **macroblock** (memories, and FUs) based placement
- Delay is improved
  - Segmented bus favors delay: bus segments have less capacitance compared to original shared bus
  - Smaller memories of first memory layer which needs to operate faster get also assigned shorter, hence faster connections

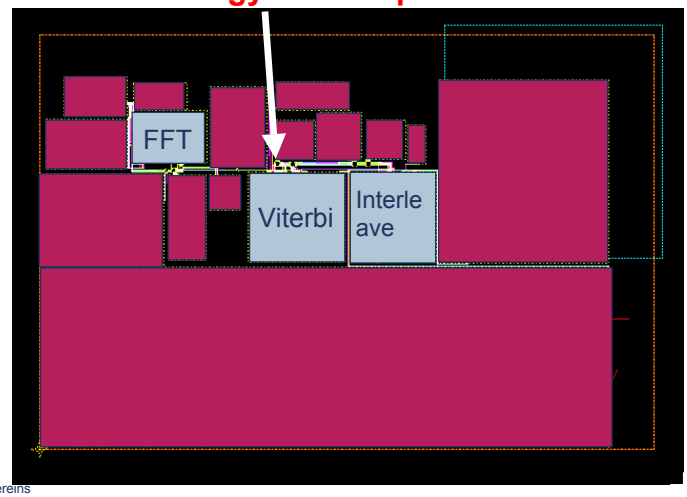
Impact of optimizations on L1 bus critical path



Bus segmentation has a significant beneficial impact on delay. Changes in the floorplanning have a minor impact.

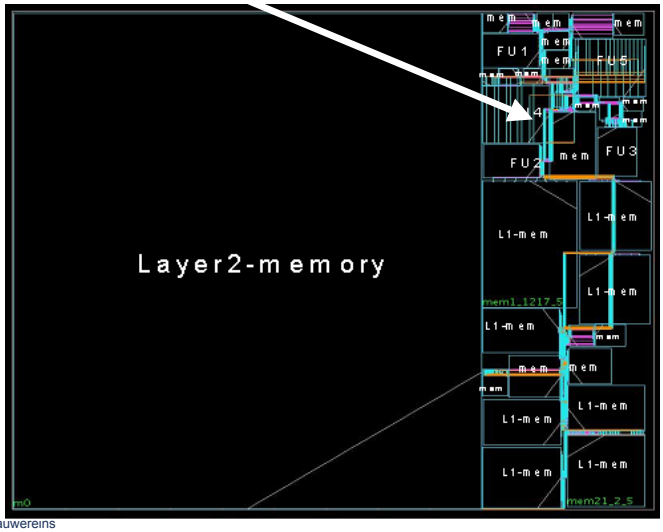
Results demonstrated on the DAB

Segmented communication network and Power aware P & R ➔ **factor 3 less comm network energy consumption**



Results demonstrated on a MPEG-4 encoder

factor 4.2 less communication network energy consumption



Conclusions

Bus segmentation: an efficient bus strategy for energy efficient memory interconnect and an enabling step for energy efficient physical design:

- Activity does not spread over entire bus: energy efficiency close to point-to-point connections'
- Average switching capacitance much smaller than in shared busses

Initial energy-aware (physical) design flow for memories also extendable to other macro-blocks:

- Segmented bus architecture
- Activity-aware (application specific) macro-block placement
- Huge gains in energy can be achieved: i.e. x3 for DAB
- Negligible area overhead (< 5%)
- Decreased delay: i.e. /3 for DAB L1 memory

Data Transfer & Storage optimizations still crucial as first step (otherwise buses of large memories will dominate)

SEEDS FOR TOMORROW'S WORLD

Rudy.Lauwereins@imec.be

IMECNOLOGY

www.imec.be

Worldwide collaboration with more than 500 companies and institutes.

IMEC – Kapeldreef 75 – B-3001 Leuven – Belgium – Tel. +32 16 281211 – Fax +32 16 229400 – www.imec.be