Bridging the gap between semiconductor technology and design:

a memory case study

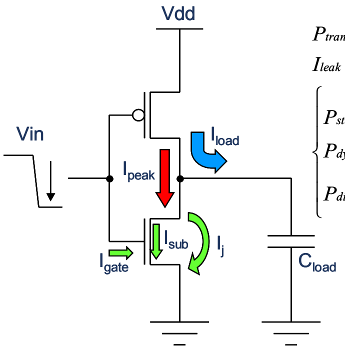
Rudy Lauwereins

Vice-President IMEC, Belgium

Professor at Katholieke Universiteit Leuven, Belgium

… beyond 100nm many technology issues become increasingly important

Transistor leakage current increase power consumption

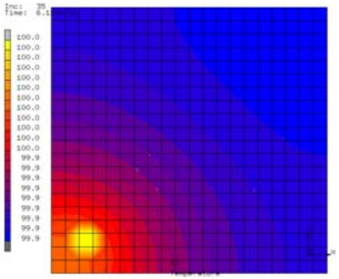


$$P_{transistor} = P_{static} + P_{dynamic} + P_{directpath}$$

$$I_{leak} = I_{gate} + I_{sub} + I_j$$

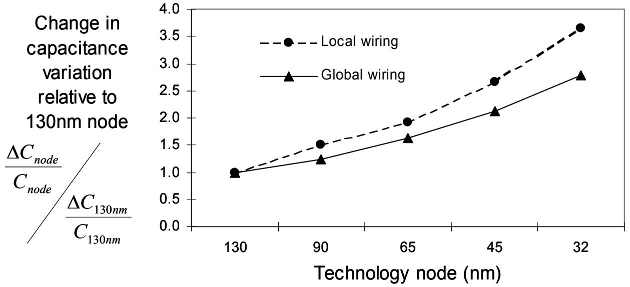
$$\begin{cases} P_{static} = I_{leak} \cdot V_{dd} \\ P_{dynamic} = C_{load} \cdot V_{dd}^2 \cdot f \\ P_{directpath} = \dfrac{t_r + t_f}{2} \cdot V_{dd} \cdot I_{peak} \cdot f \approx 0 \end{cases}$$

Temperature driven dynamic process variations



Increased static process variations with scaling

- Relative spread of capacitance due to technology

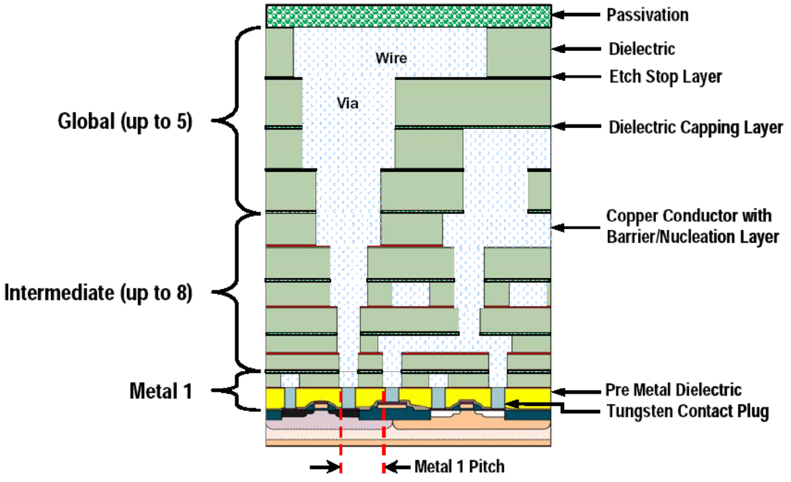


⇒Relative spread in capacitance compared to 130 nm technology node (for dense wiring): increase of factor 2 in 65 nm node, factor 3.5 in 32 nm node

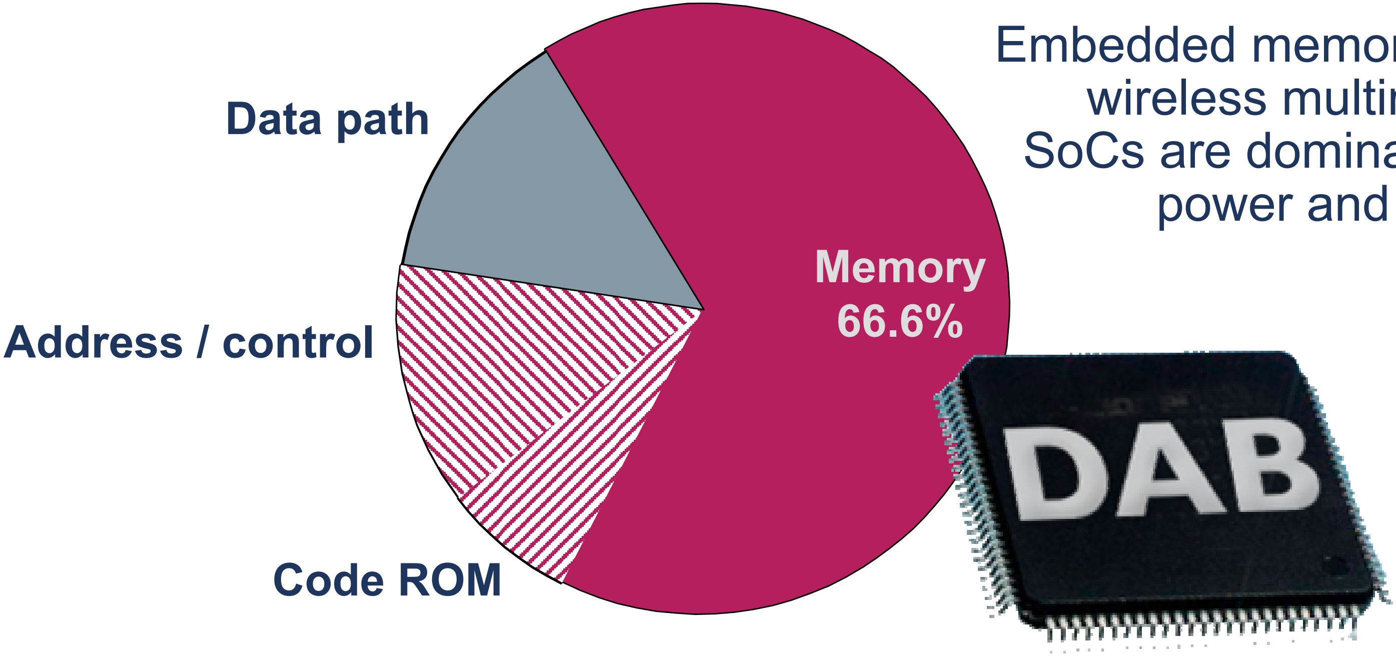
⇒This is including the effect of CMP and roughness variation

Capacitance and resistivity of local wires increases with scaling

Global wires do not scale in length



Memories dominate power consumption in data-dominated applications



Embedded memories in wireless multimedia SoCs are dominant for power and delay

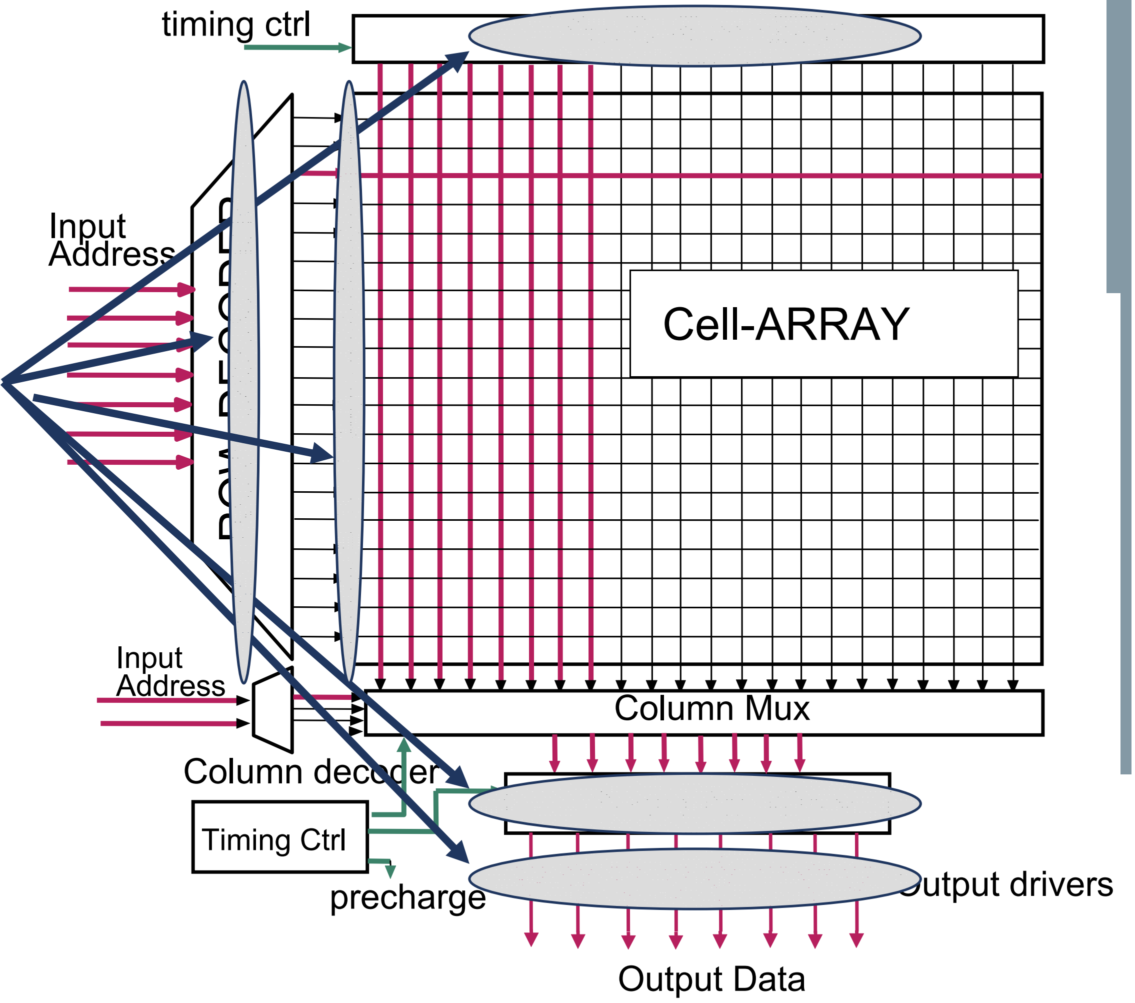
The current focus of the TAD program is on SRAMs:

- SRAM is key element (energy and delay) in system (stand alone or embedded)
- Easy to model because regular and predictable topology (standard cell design with Place&Route in the flow is stochastic)
- Advantage for critical lithography and as technology driver (Layout rules are typically smaller than rest of components)

Key element in new approach: “knobs” in memory to create Pareto trade-offs

Buffers/drivers needed/present in different parts of memory architecture

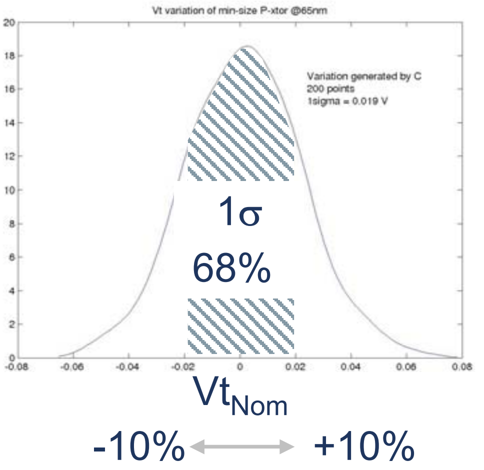
Buffers



Limited impact in area but big impact in energy/delay

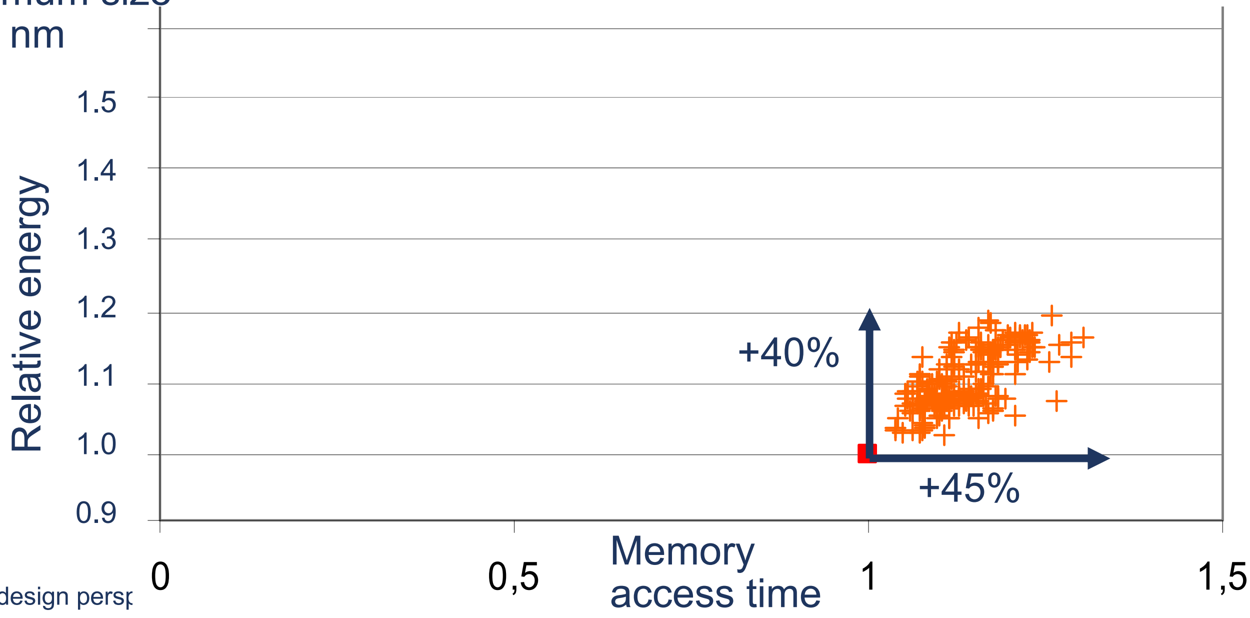
Ideal components for cheap Pareto “Config knobs”

IMEC's concept combines 100% (parametric) yield with variation tolerance

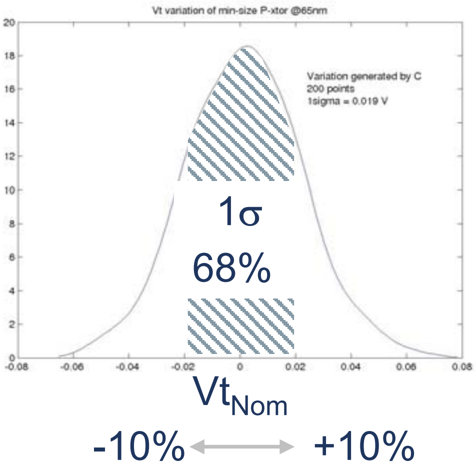


Variation for minimum size TOR @ 65 nm

A moderate 10% variation for one transistor leads to 40% variation in access time for a 1KB memory



IMEC's concept combines 100% (parametric) yield with variation tolerance

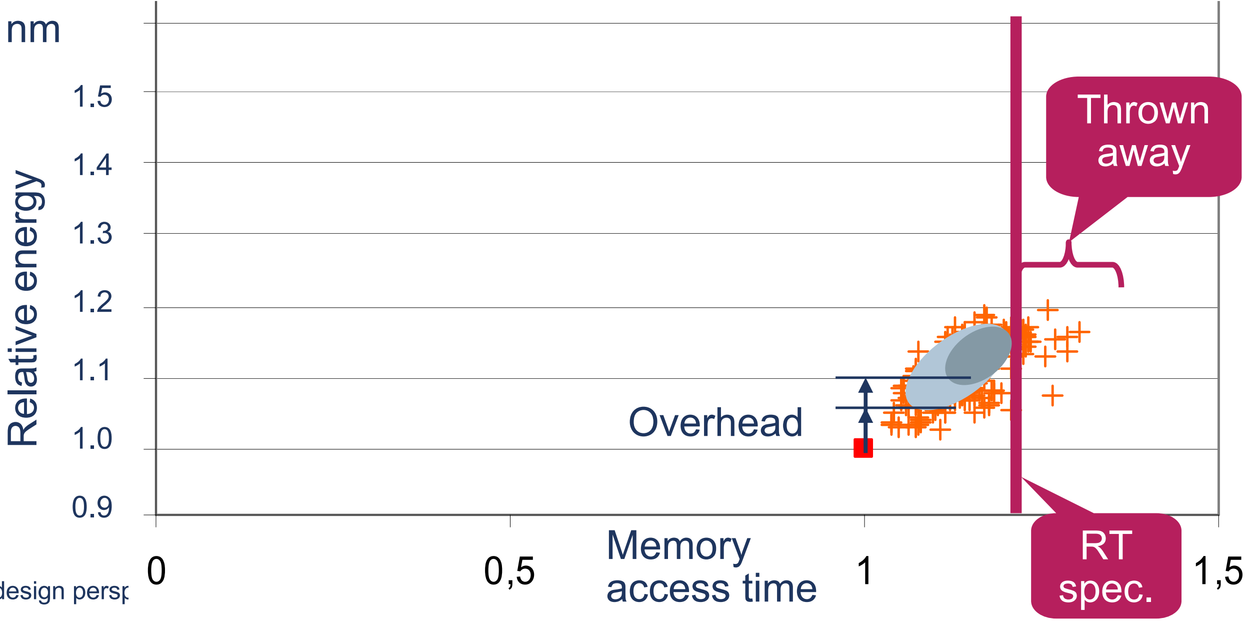


Variation for minimum size TOR @ 65 nm

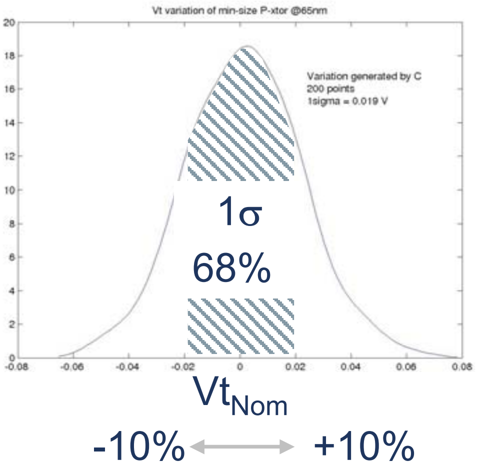
A moderate 10% variation for one transistor leads to 40% variation in access time for a 1KB memory

Sigma-based design improves (parametric) yield at the cost of performance-power overhead (design margins)

Sigma-based design badly scales: new silicon nodes have higher variation and hence more overhead needed



IMEC’s concept combines 100% yield with variation tolerance



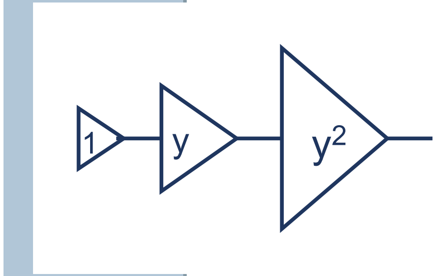
Variation for minimum size TOR @ 65 nm

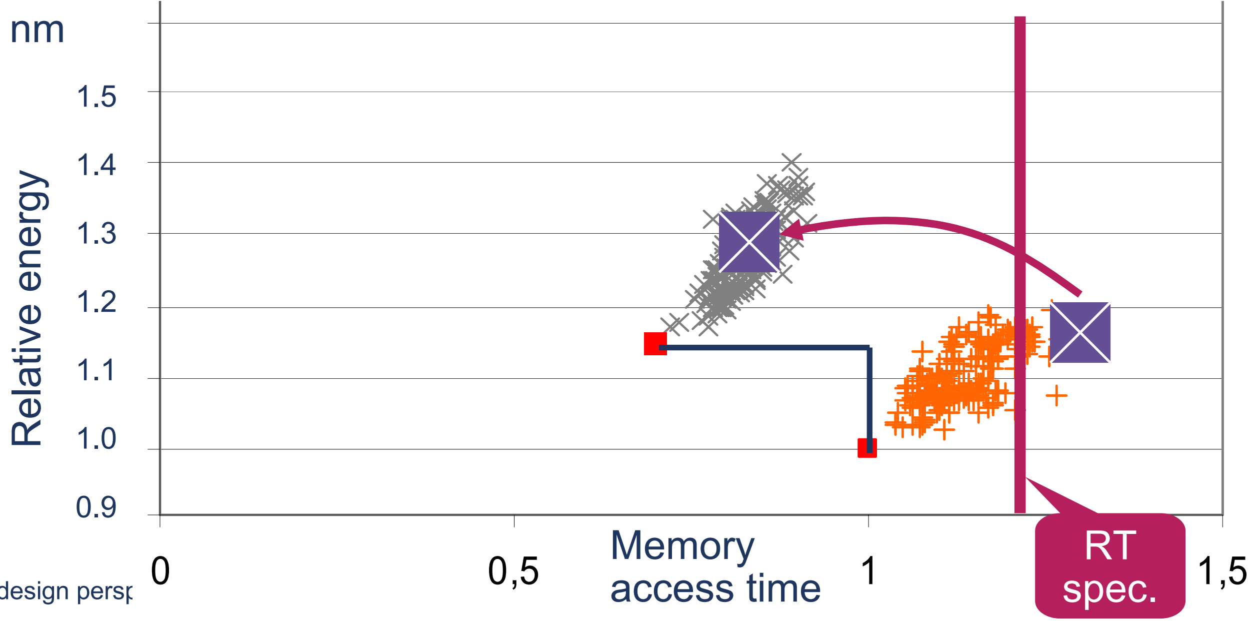
A moderate 10% variation for one transistor leads to 40% variation in access time for a 1KB memory

Sigma-based design leads to low yield and overhead

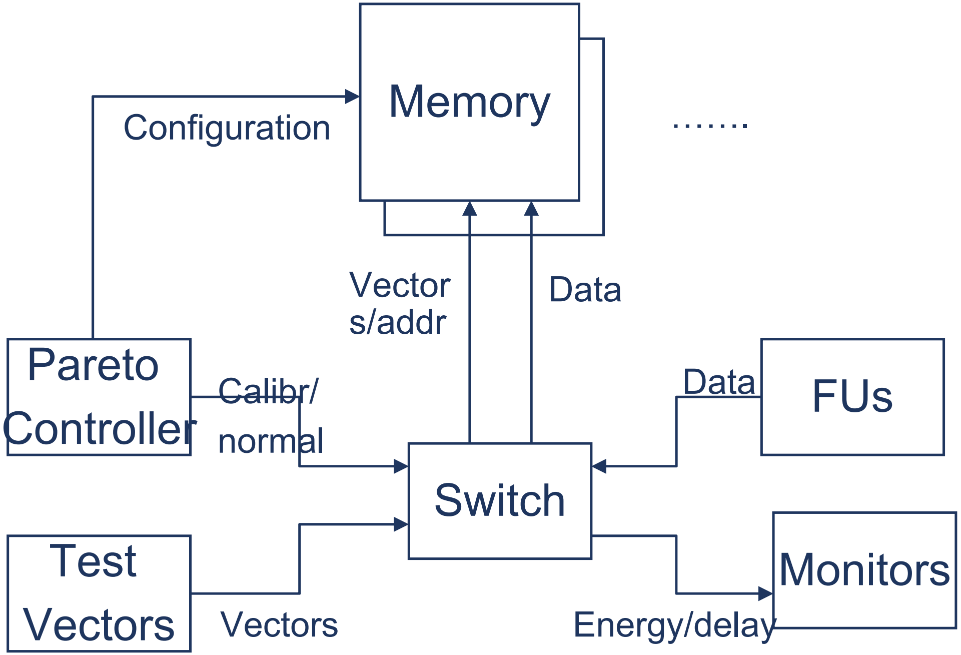
Sigma-based design badly scales: new silicon nodes have higher variation and hence more overhead

Better solution: just live with the speed you get & for those falling outside the spec, switch to a different/faster implementation, e.g. by using an additional driver step





System requirements: run-time Pareto controller and calibration loop



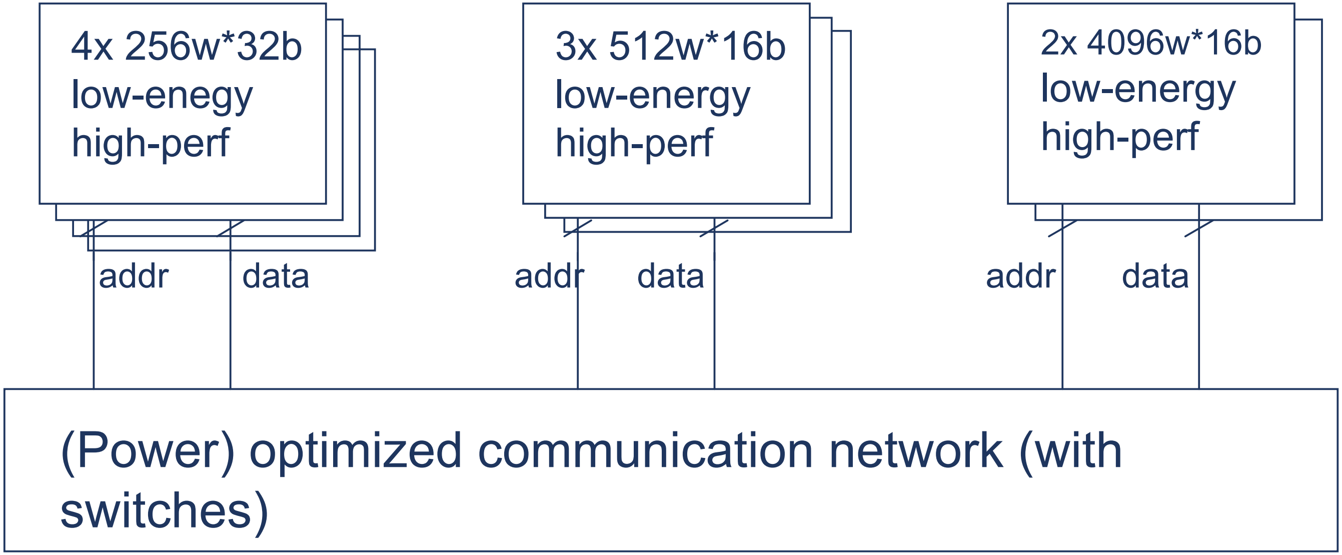
Calibration (rarely): Per memory:

1. Apply Test vectors
2. Measure E/D
3. Overwrite Pareto tables

Normal operation:

1. Determine Pareto operating point for all mems in Pareto controller
2. Steer configuration knobs in memories

Approach is applied to the memory organisation in a DAB application



4x 256w*32b
low-enegy
high-perf

3x 512w*16b
low-energy
high-perf

2x 4096w*16b
low-energy
high-perf

addr

data

addr

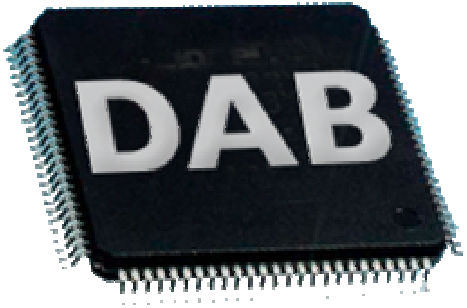
data

addr

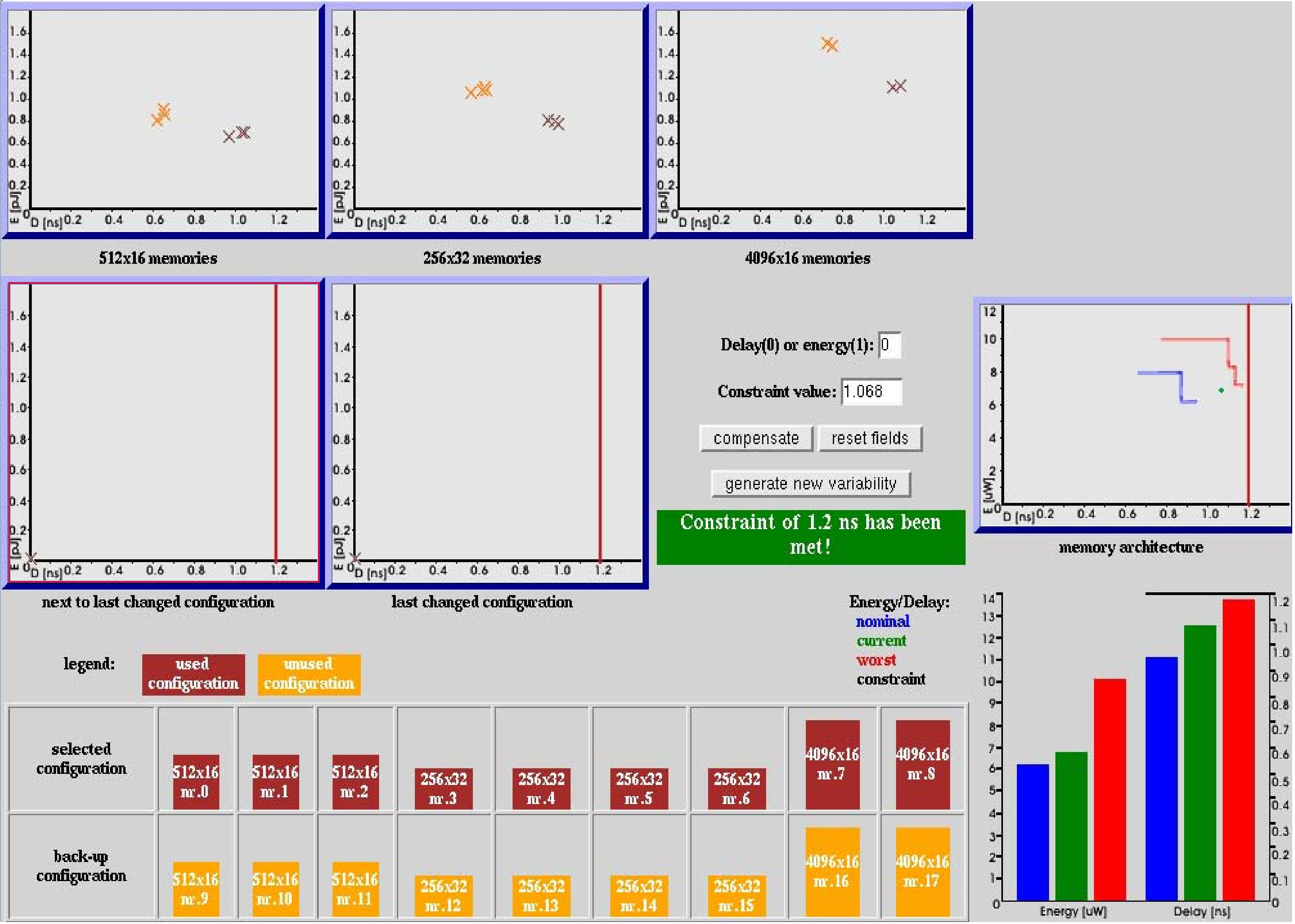
data

(Power) optimized communication network (with switches)

Base Implementation: 7 x 1KByte (16+32bit)+ 2 x 8KByte SRAMs with two configurations “knobs” each for Pareto trade-offs (low-energy and high-performance)



DEMO - DAB: illustration of system level adaptation to process variability



Delay(0) or energy(1): 0

Constraint value: 1.068

Constraint of 1.2 ns has been met!

legend: used configuration | unused configuration

selected configuration	512x16 nr.0	512x16 nr.1	512x16 nr.2	256x32 nr.3	256x32 nr.4	256x32 nr.5	256x32 nr.6	4096x16 nr.7	4096x16 nr.8
back-up configuration	512x16 nr.9	512x16 nr.10	512x16 nr.11	256x32 nr.12	256x32 nr.13	256x32 nr.14	256x32 nr.15	4096x16 nr.16	4096x16 nr.17

Conclusions

Problem:

- Small on-chip SRAM is critical component (L1-memories)
- Impact process variability at SRAM level much more dramatic than transistor (from 10% to 50%)
- Industry → sigma-based design minimizes variability but trade-off yield and generates overhead (critical for L1-memories)

Alternative:

- Use best case SRAM design tolerating variability with 100% yield (functionality still tested)
- Provide configuration "knobs" offering wide range of energy/delay trade-offs
- Let system compensate for eventual drift in variability at architecture level (system timing and not clock cycle based)

Feasibility:

- Concept demonstrated in DAB receiver at SPICE level imec

SEEDS FOR TOMORROW'S WORLD **IMEC***NOLOGY*