Dinesh Pamunuwa*, Matt Grange**, Axel Jantsch+,
Sunil Rana*, Tyson Tian Qin*

*University of Bristol, Bristol UK

**Mentor Graphics, USA

+KTH Royal Institute of Technology, Stockholm, Sweden

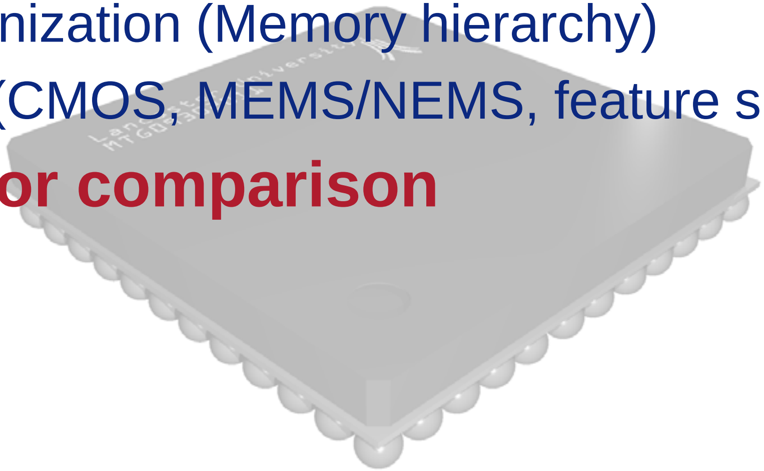
System Performance Analysis for Heterogeneous 3-D ICs and Emerging Technologies

University of BRISTOL

microelectronics research group

Future Direction of 3-D ICs

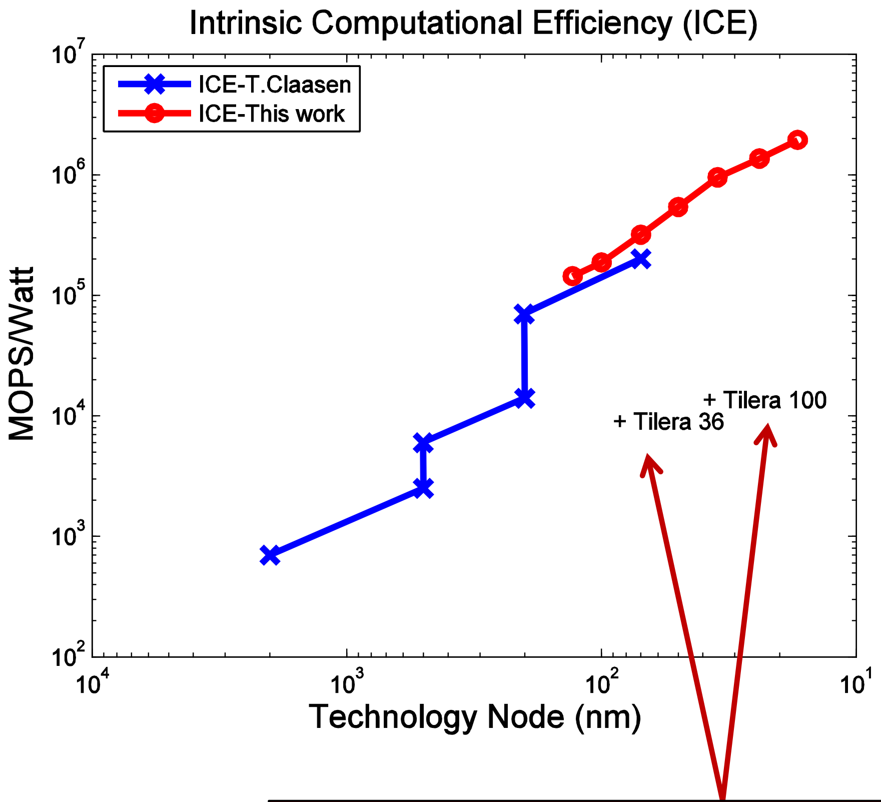
- ❑ **3-D ICs promise performance / cost advantages for high performance digital applications as well as enhanced functionality**
- ❑ **A staggering amount of degrees of freedom exist**
  - – Packages (2-D, Multi-chip Modules, 3-D die stacks)
  - – Digital architecture (single to multi to many-core)
  - – Routing architectures (buses, Networks-on-Chip, hybrids)
  - – System organization (Memory hierarchy)
  - – Technology (CMOS, MEMS/NEMS, feature size reduction)



- ❑ **Framework for comparison**

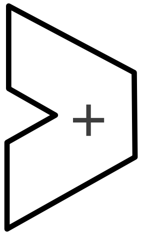
Intrinsic Computational Efficiency

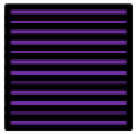
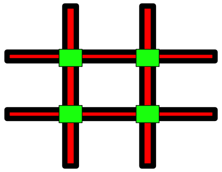
- **The Intrinsic Computational Efficiency (ICE), proposed by T. Claasen, creates the maximum upper bound for computational capability of a silicon-based processor.**
  - The entire silicon area of a processor is filled with the most fundamental computational unit, in this case we have used 32-bit adders.
  - A real system could never achieve the same performance per Watt because this metric ignores the overhead of control circuitry, interconnect, and memory.



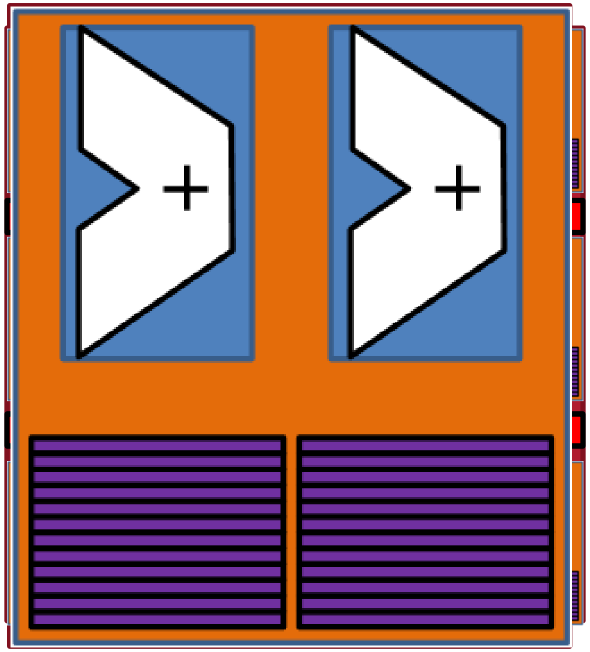
Two recent multi-core processors

Expanding the *ICE* to the *ECE*

- ❑ The *ICE* gives the ***maximum upper bound*** on efficiency, but cannot account for realistic systems because it only considers the computational unit.
- ❑ We build upon the *ICE* by modelling the three fundamental operations of any processing unit:
  - The computational operation 

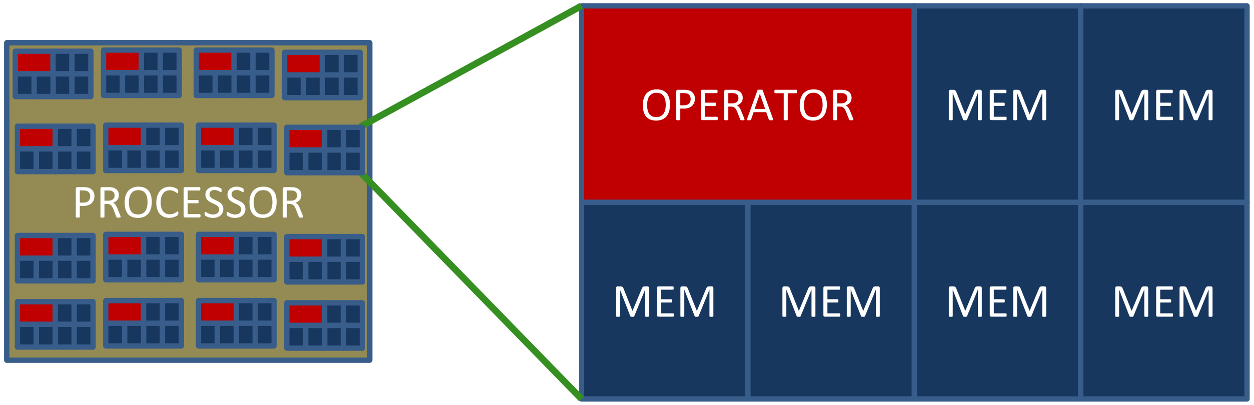
  - The memory 
  - The interconnect 

Processing Core



Temporal and Spatial Organization of Memory

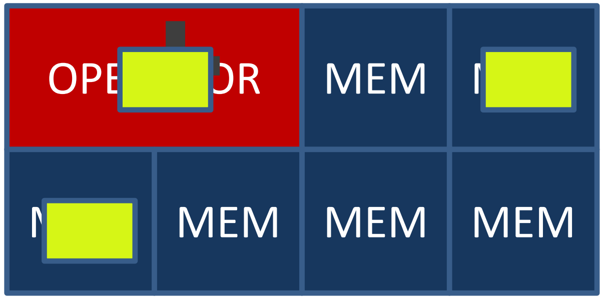
- ❑ μ_s
  - – Gives us the amount of memory per operator. Think of it as the amount of on-chip cache available



- ❑ μ_T
  - – Gives us the number of memory reads/writes per operation.

2 reads, 1 write

$\mu_T = 3$

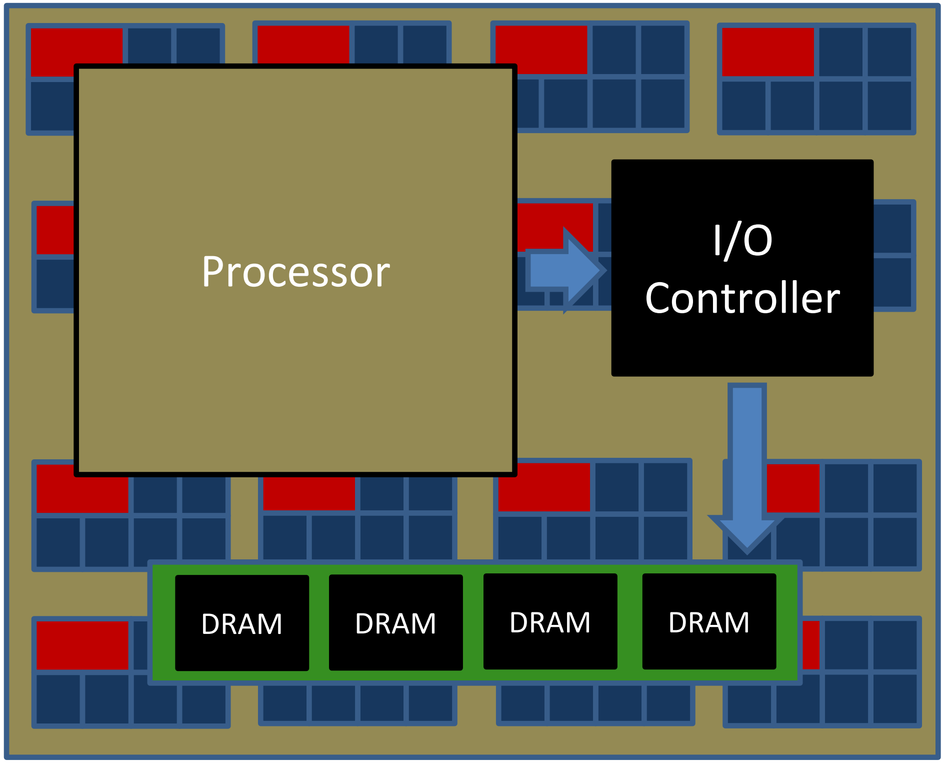


On- or Off-chip Memory?

- ω (0-1)
  - Gives the ratio of on- to off-chip memory in the system. Off-chip memory requires exiting the die with I/O drivers to external chips.

0: all-off-chip

1: all-on-chip



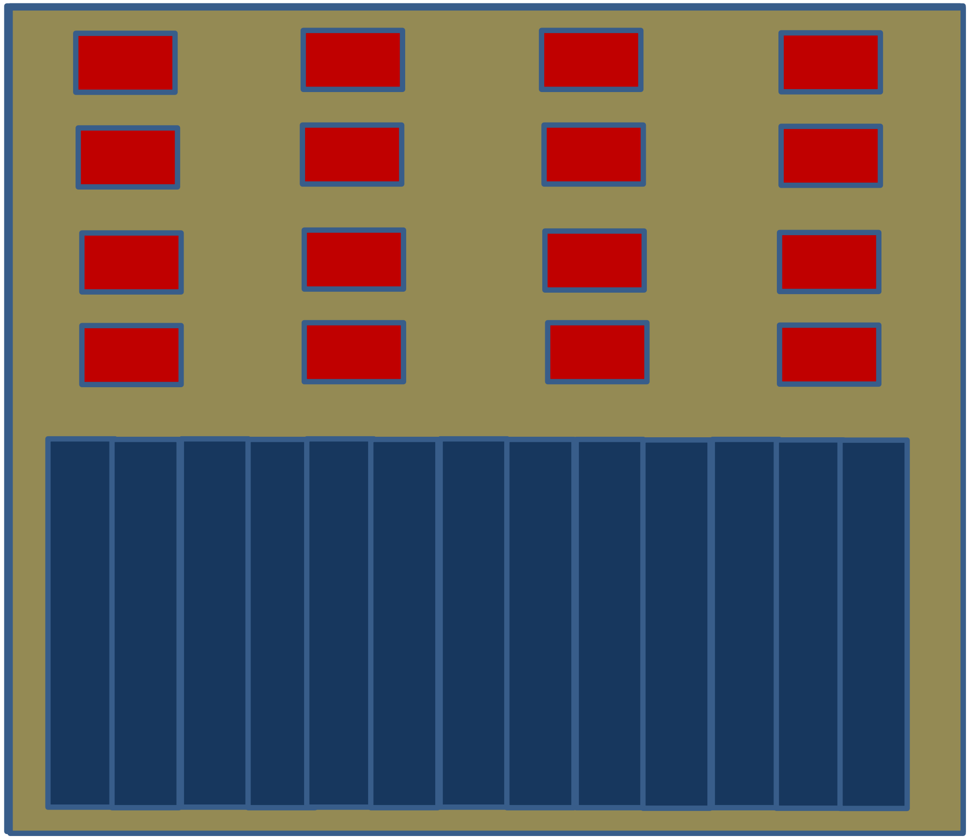
Memory Distribution Factor

- ❑ Δ (0-1)
  - Gives the distribution factor of the memory, or how close (local) it is the operator.

0:all local

0.5:semi-local

1:non-local



Effective Computational Efficiency

- ❑ **For each computation we must consider the expense of energy for the operation, interconnect (on and off-chip) and memory reads/writes.**

Number of Memory Accesses/op

Energy for a 32-bit addition

On-chip memory energy

Off-chip memory energy

$$EE^{tn}_{arch} = \boxed{E^{tn}_{32}} + \boxed{\mu_T}\left(\boxed{\omega}\boxed{\left(e_1 + \Delta \times E\_int^{tn}_{arch}\right)} + \boxed{(1-\omega)}\boxed{\left(e_1 + E\_int^{tn}_{arch} + E_{offchip}\right)}\right)$$

Ratio of on-chip : off-chip memory

$$ECE^{tn}_{arch} = \boxed{\frac{1}{EE^{tn}_{arch}}}$$

Amount of computation possible within the envelope of 1 Joule

Modelling Hierarchy

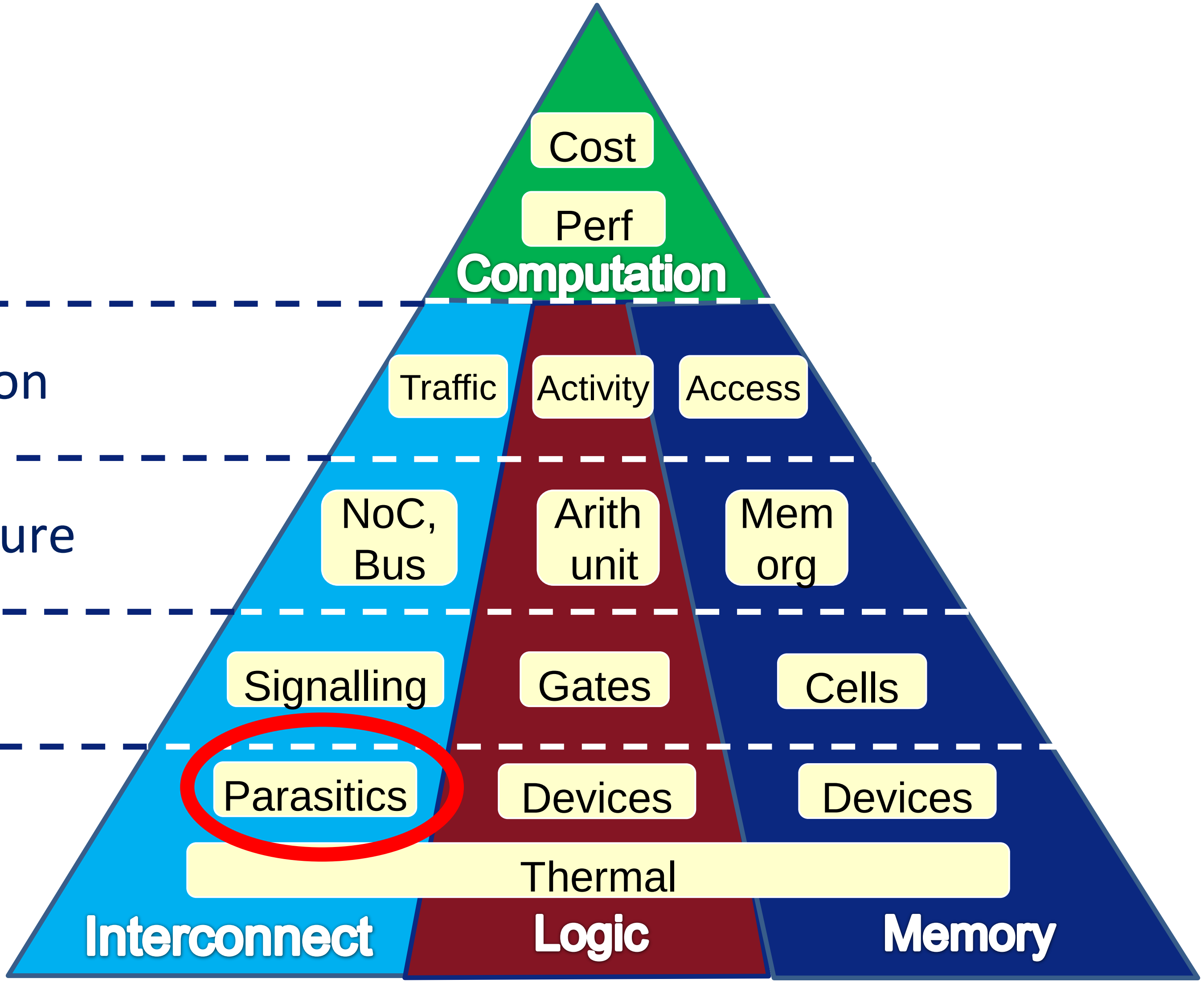
System

Application

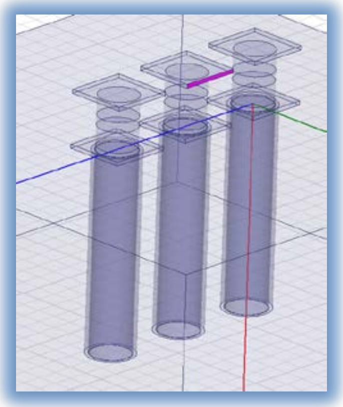
Architecture

Circuit

Physical



Model Development

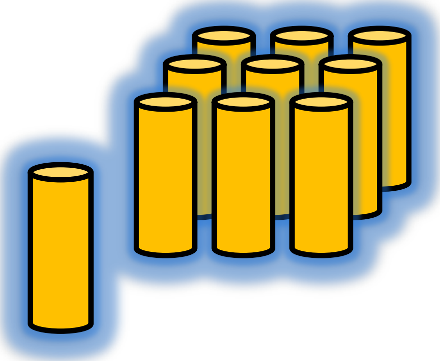


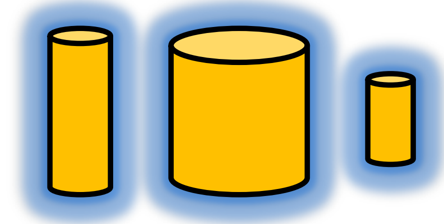
Parasitic Extraction (Q3D)

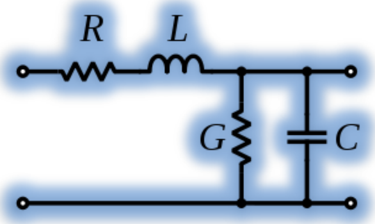
*Resistance (**R**), Inductance (**L**)*
*Conductance (**G**), Capacitance (**C**)*

Geometrical Sweeps (length, radius, pitch, etc.)

Physical Sweeps (Topology, Substrate, etc.)





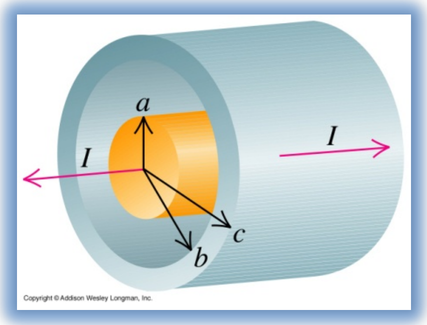


Data Manipulation (MATLAB)

Modelling

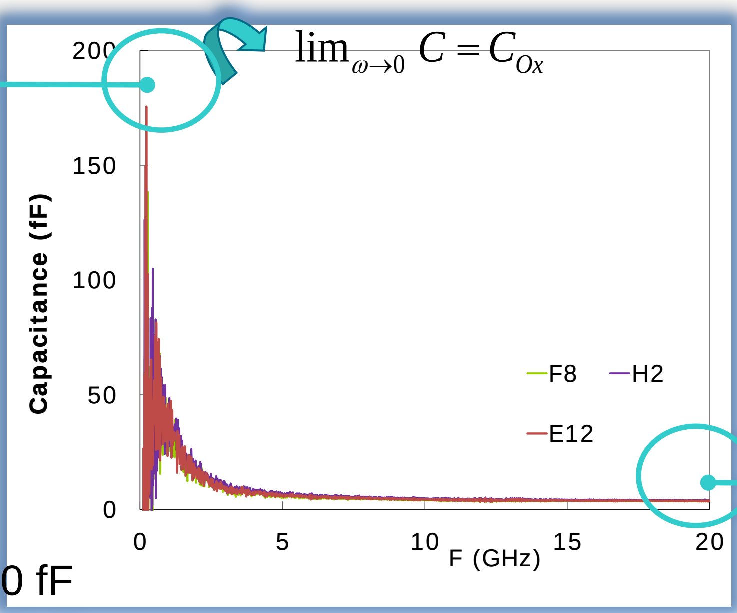
Physical Principles

Verification



Copyright © Addison Wesley Longman, Inc.

Model Verification



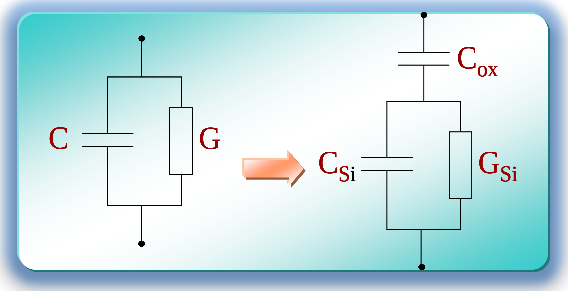
$$\lim_{\omega\to 0} C = C_{Ox}$$

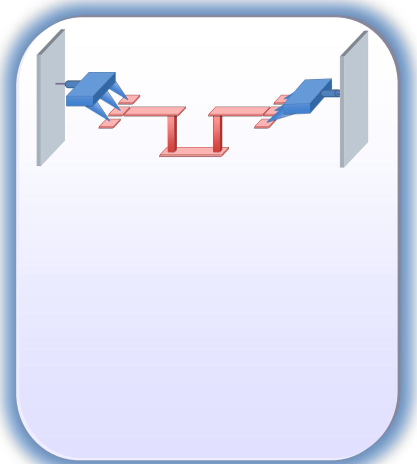
320 fF

(306.6)

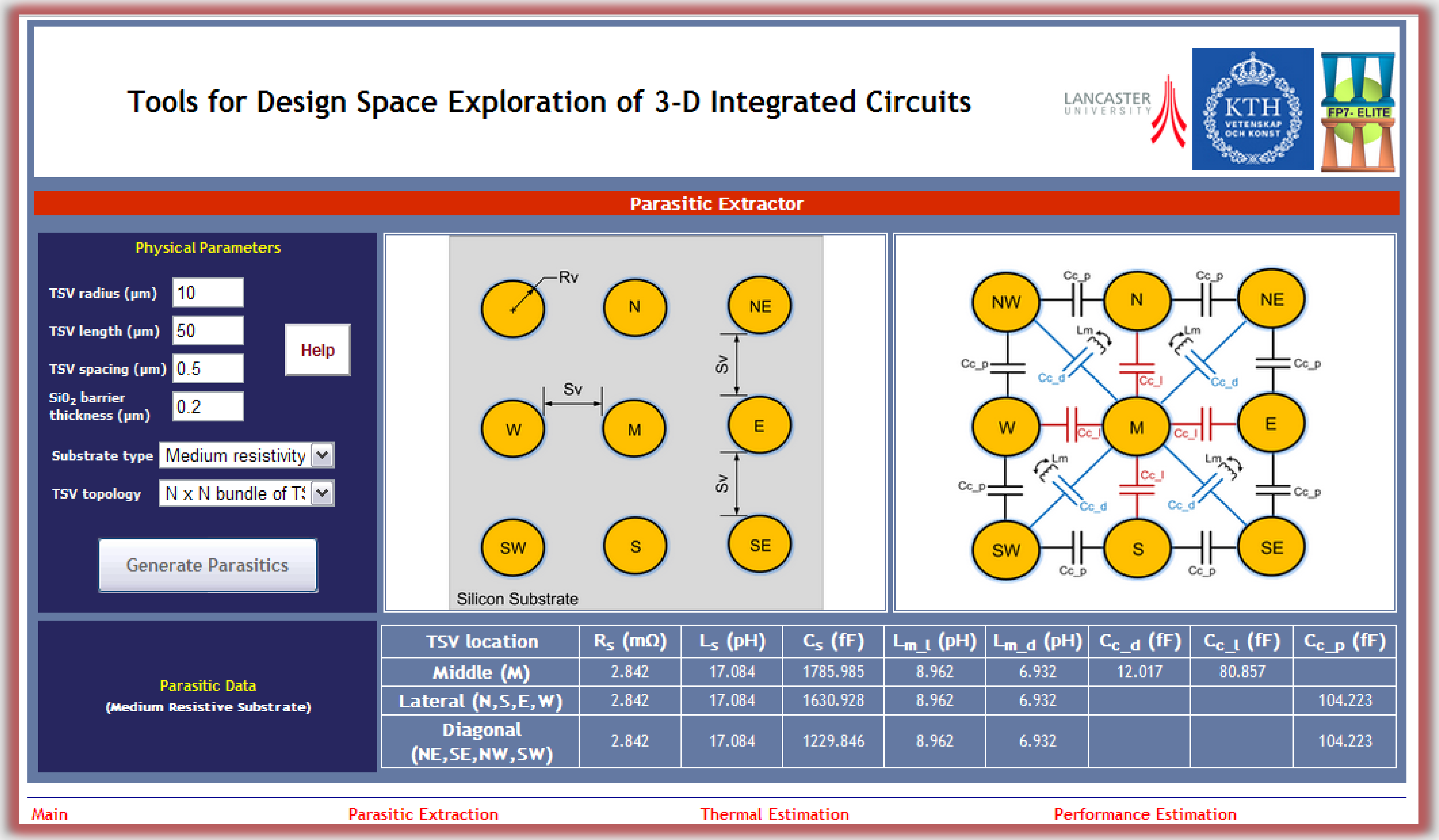
4 fF

(4.5)





Stand-alone Parasitic Estimation Tool



Tools for Design Space Exploration of 3-D Integrated Circuits

Parasitic Extractor

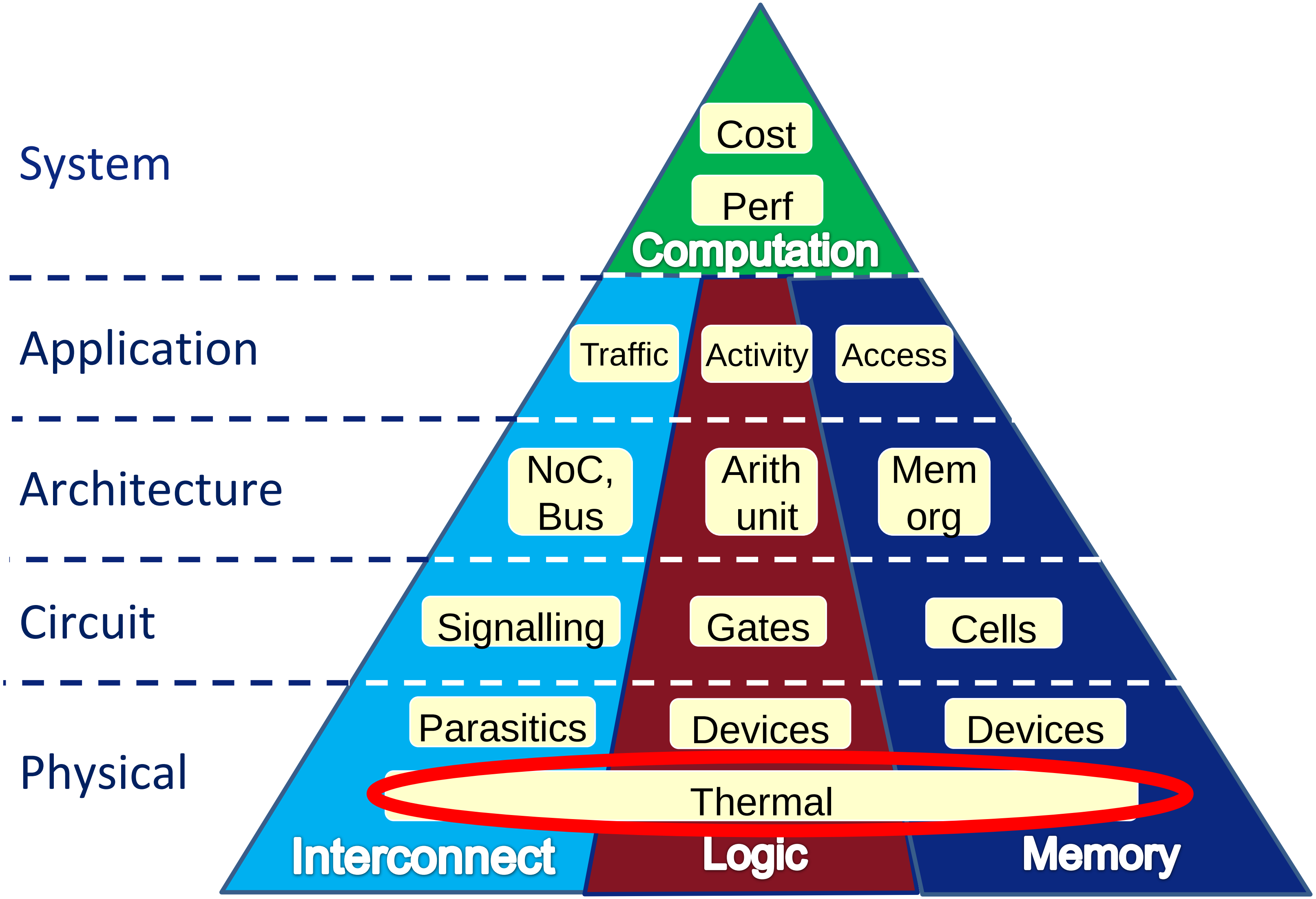
Physical Parameters

TSV radius (μm)	10
TSV length (μm)	50
TSV spacing (μm)	0.5
SiO_2 barrier thickness (μm)	0.2
Substrate type	Medium resistivity
TSV topology	N x N bundle of TS

Parasitic Data
(Medium Resistive Substrate)

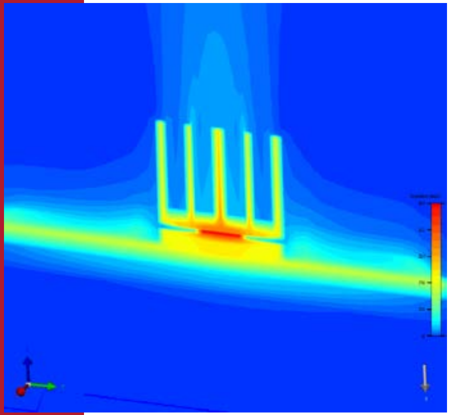
TSV location	R_S (mΩ)	L_S (pH)	C_S (fF)	L_{m_l} (pH)	L_{m_d} (pH)	C_{c_d} (fF)	C_{c_l} (fF)	C_{c_p} (fF)
Middle (M)	2.842	17.084	1785.985	8.962	6.932	12.017	80.857	
Lateral (N,S,E,W)	2.842	17.084	1630.928	8.962	6.932			104.223
Diagonal (NE,SE,NW,SW)	2.842	17.084	1229.846	8.962	6.932			104.223

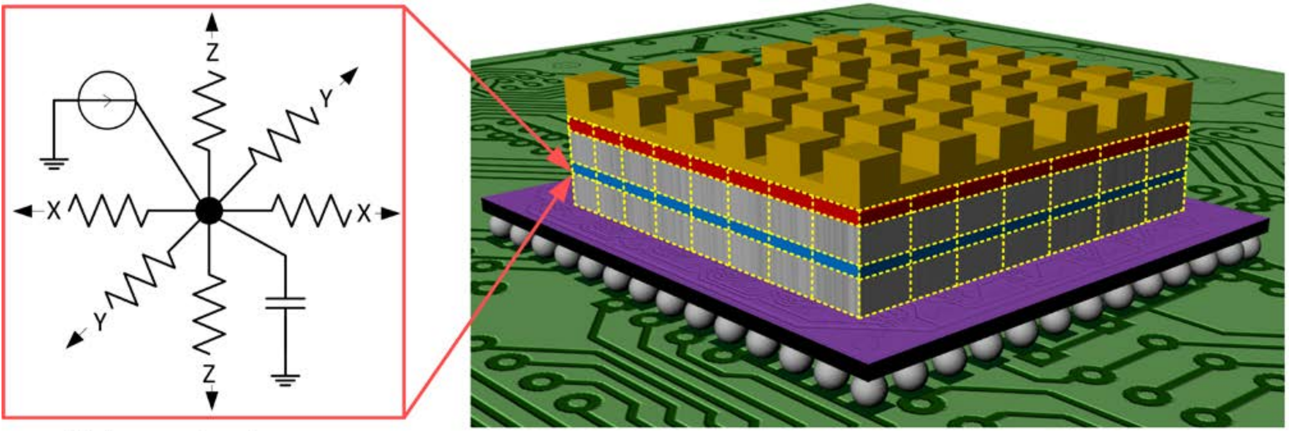
Modelling Hierarchy



Thermal Behaviour

- **Compact thermal models have been developed as part of the toolset to quickly predict the thermal behaviour.**
  - Verification based on comparison with results from a Computational Fluid Dynamic (CFD) solver (FloTHERM).
  - The thermal behaviour and limitations of 2-D and 3-D packages has been extracted from simulations.





● Voltage at junction equivalent to temperature

Material Thermal Resistance

Material Thermal Absorption Capability (specific heat)

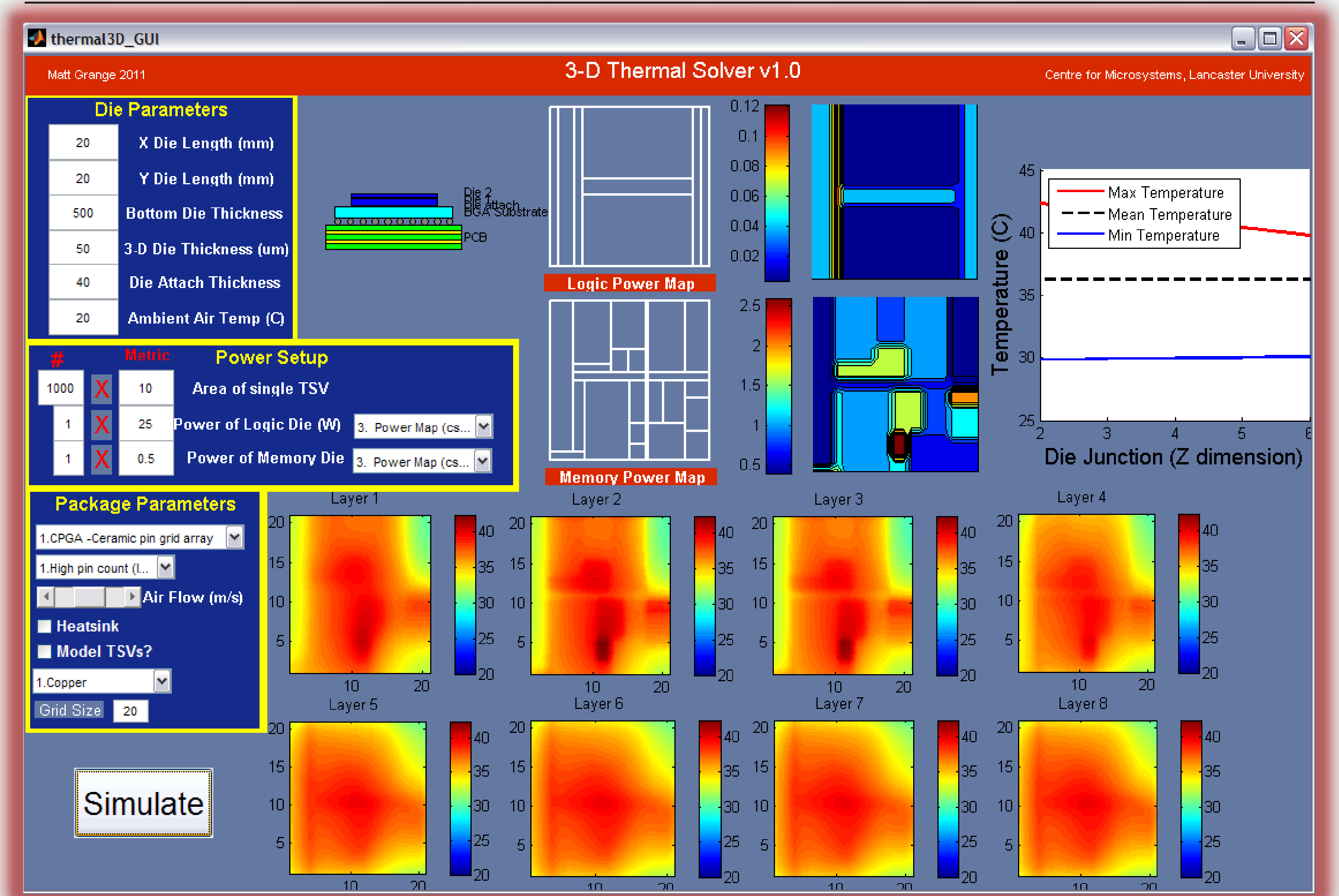
Power Injected into junction

$$R_{th} = \frac{L}{kA}$$

$$T_j = qR_{th} + T_a$$

$$\theta_{j-a} = \frac{T_j - T_a}{q}$$

Stand-alone 3-D Thermal Tool



thermal3D_GUI

Matt Grange 2011

3-D Thermal Solver v1.0

Centre for Microsystems, Lancaster University

Die Parameters

20	X Die Length (mm)
20	Y Die Length (mm)
500	Bottom Die Thickness
50	3-D Die Thickness (um)
40	Die Attach Thickness
20	Ambient Air Temp (C)

Power Setup

#		Metric		
1000	X	10	Area of single TSV	
1	X	25	Power of Logic Die (W)	3. Power Map (cs...
1	X	0.5	Power of Memory Die	3. Power Map (cs...

Package Parameters

1.CPGA -Ceramic pin grid array

1.High pin count (l...

Air Flow (m/s)

□ Heatsink

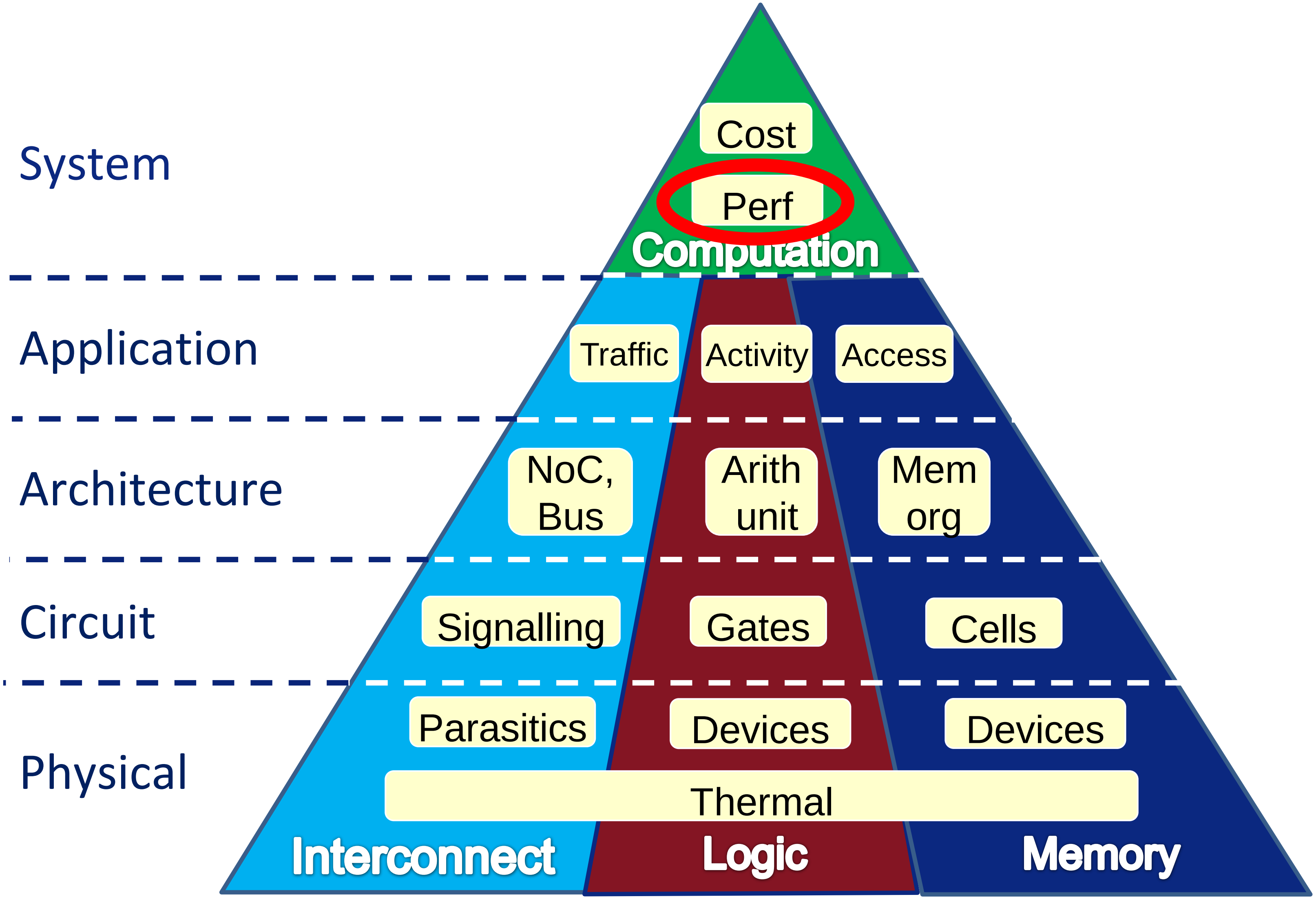
□ Model TSVs?

1.Copper

Grid Size 20

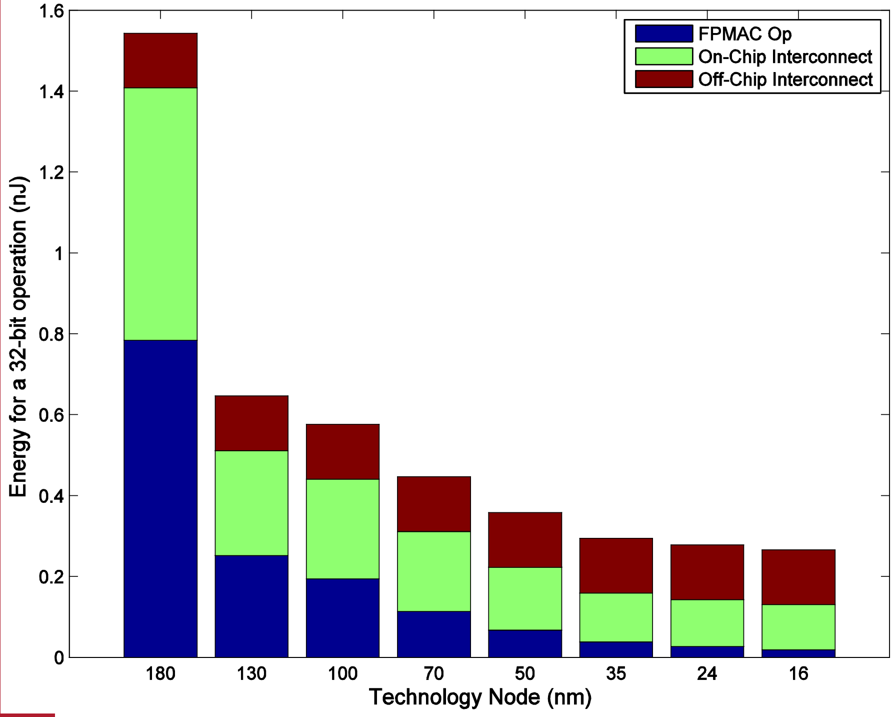
Simulate

Modelling Hierarchy

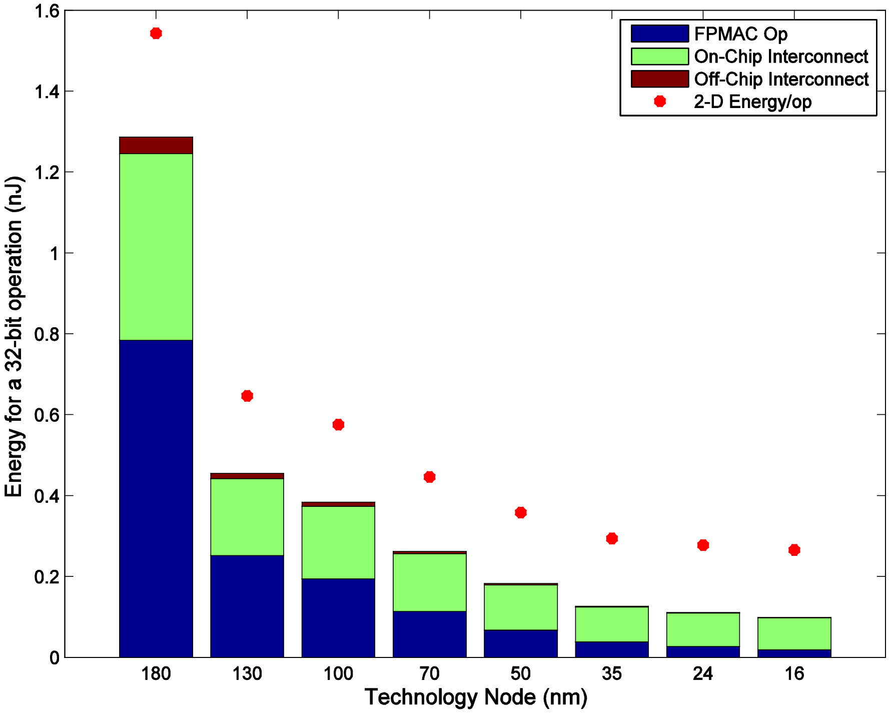


Scaling 2-D versus 3-D DRAM

2-D with off-chip DRAM



2-layer 3-D with in-stack DRAM



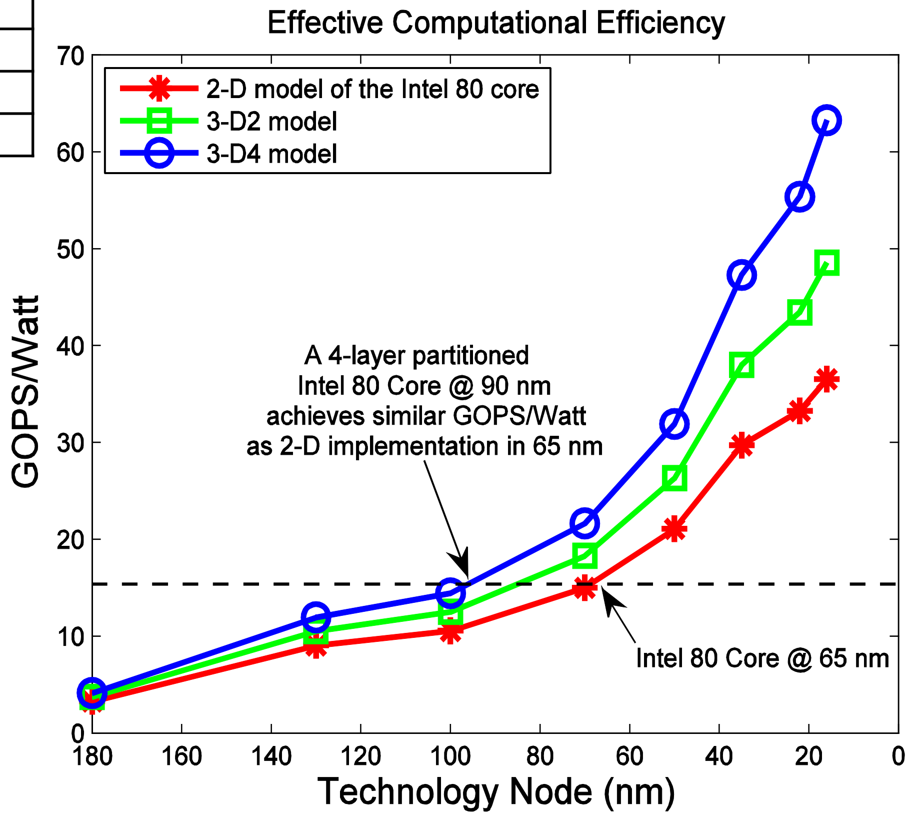
Fixed System: Intel 80 Core

Param.	Type	Intel 80 Core	Equivalent
N	# of Layers	1 (2-D)	1-16 (3-D)
A	Die Area	12.64×21.72 mm	275 mm2/N
tn	Tech. node	65 nm	180-17 nm
b	Data width	32-bit	32
μ_s	Memory/Operator	2K SRAM/2 FPU	1KB/Op
μ_t	Memory/Operation	App. Specific	01-Mar
σ	Bus Sharing Ratio	8x10 mesh/160 FPU	18/160=0.11
Δ	Memory Distribution	NoC Mesh	0.01-0.1
ω	On/off-chip mem	All on-chip	1
P	Power (W)	20-230	App. Specific

Implemented processors can be modelled by varying the parameters

The 4-layer 80-core 3-D system at 90 nm is still better than a 2-D system at 65 nm

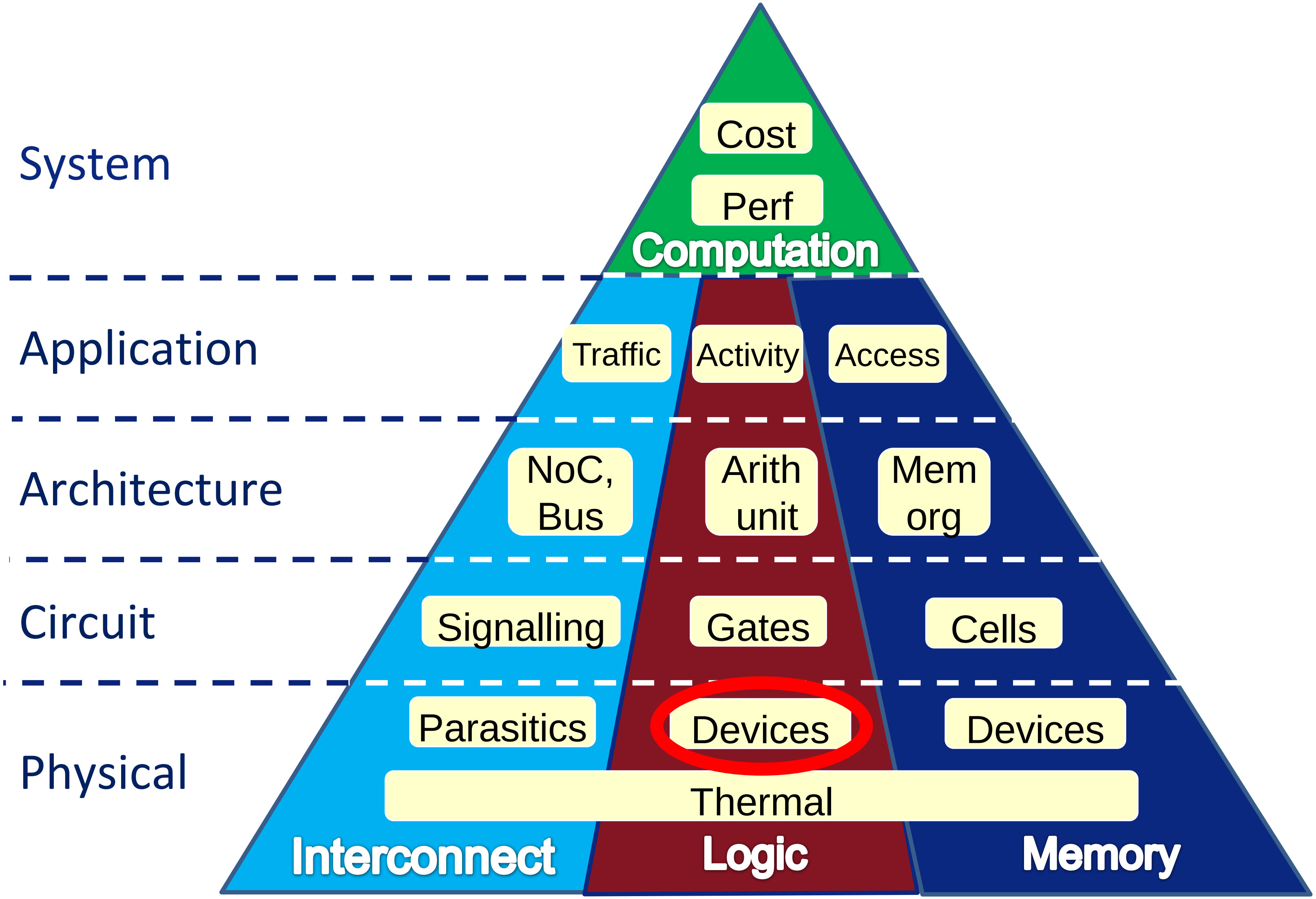
For every doubling of the stack height the computational efficiency increases by 20-30%


Effective Computational Efficiency

A 4-layer partitioned Intel 80 Core @ 90 nm achieves similar GOPS/Watt as 2-D implementation in 65 nm

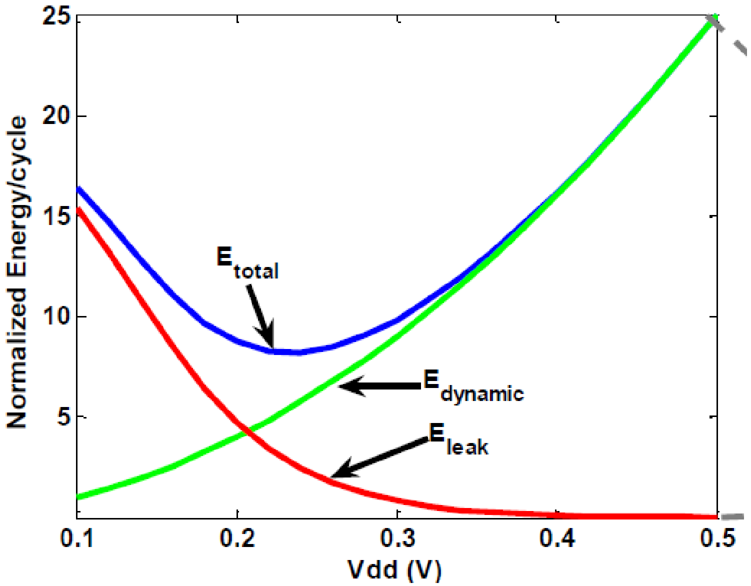
Intel 80 Core @ 65 nm

Modelling Hierarchy

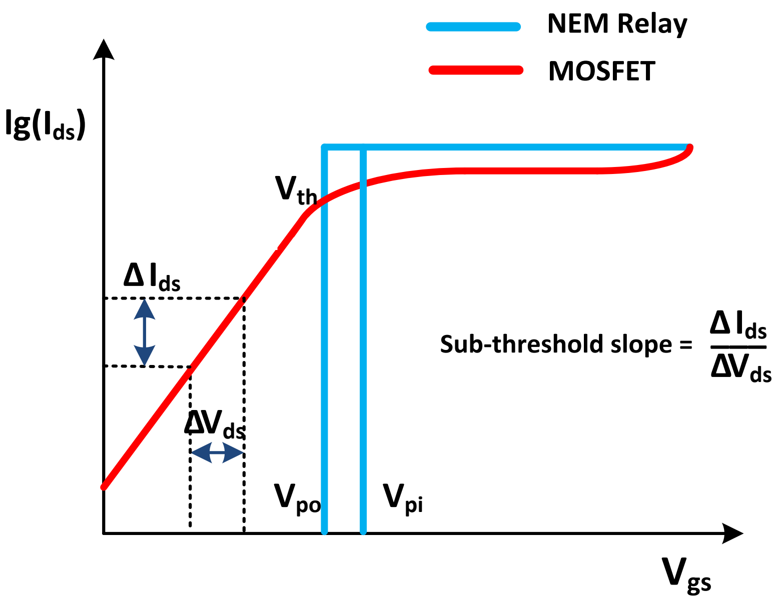


NEM Relay Based Computation

❑ Limitation of CMOS Energy Efficiency



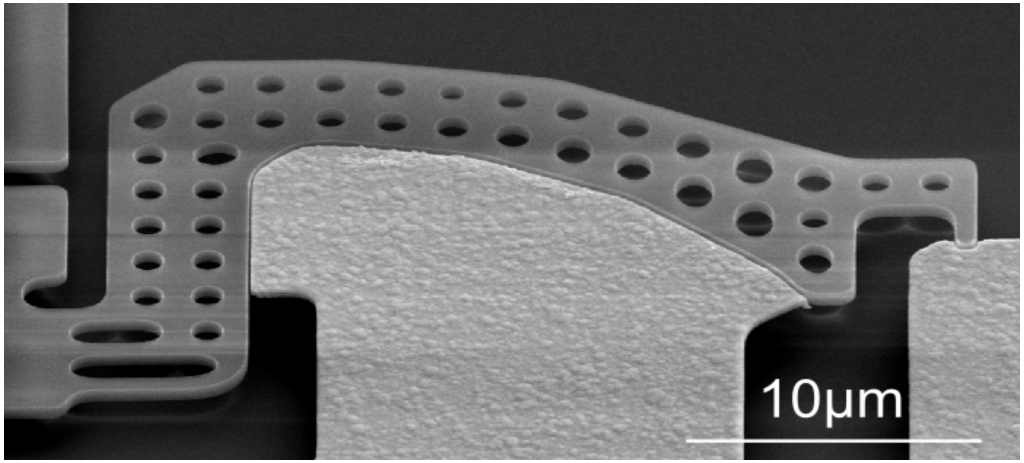
(source: F. Chen et al, ICCAD 2008)



❑ NEM relay advantages

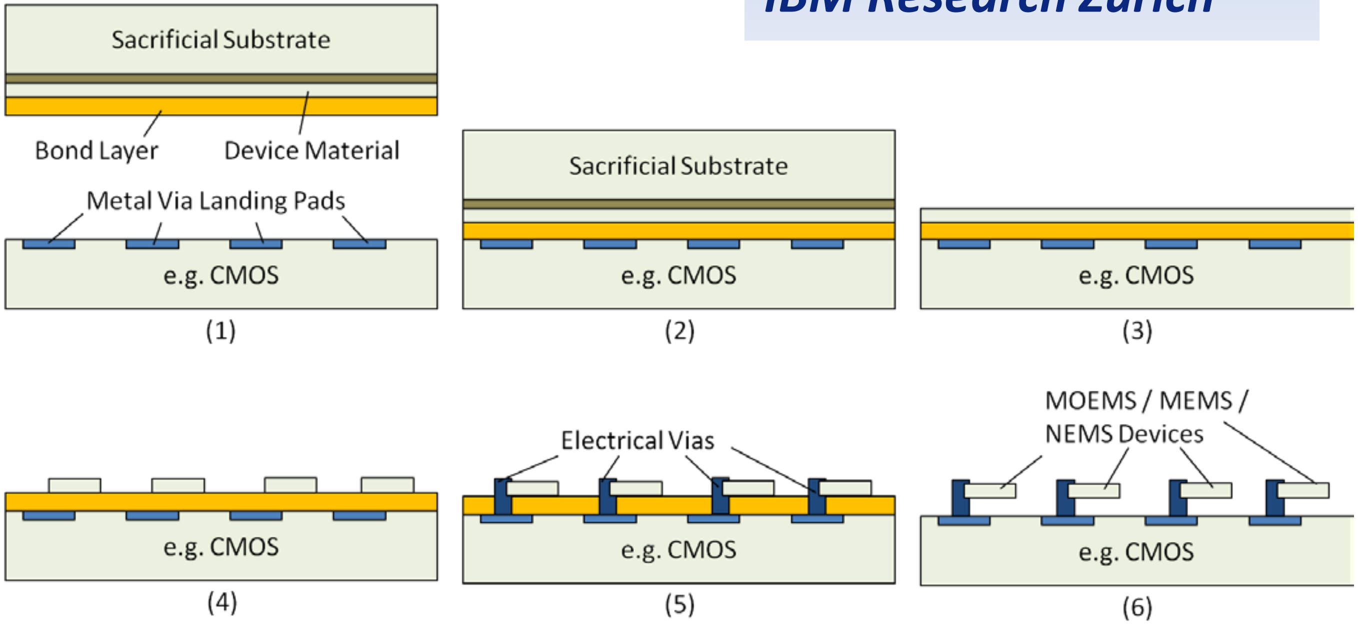
- practically zero leakage
- Very steep slope for turn-on/-off transient
- high on-current

NEM Relay Technology



In-plane switch, fabricated using standard lithography in NEMIAC project; nm gap using sacrificial layer
courtesy D. Grogg et al. IBM Research Zurich[1]

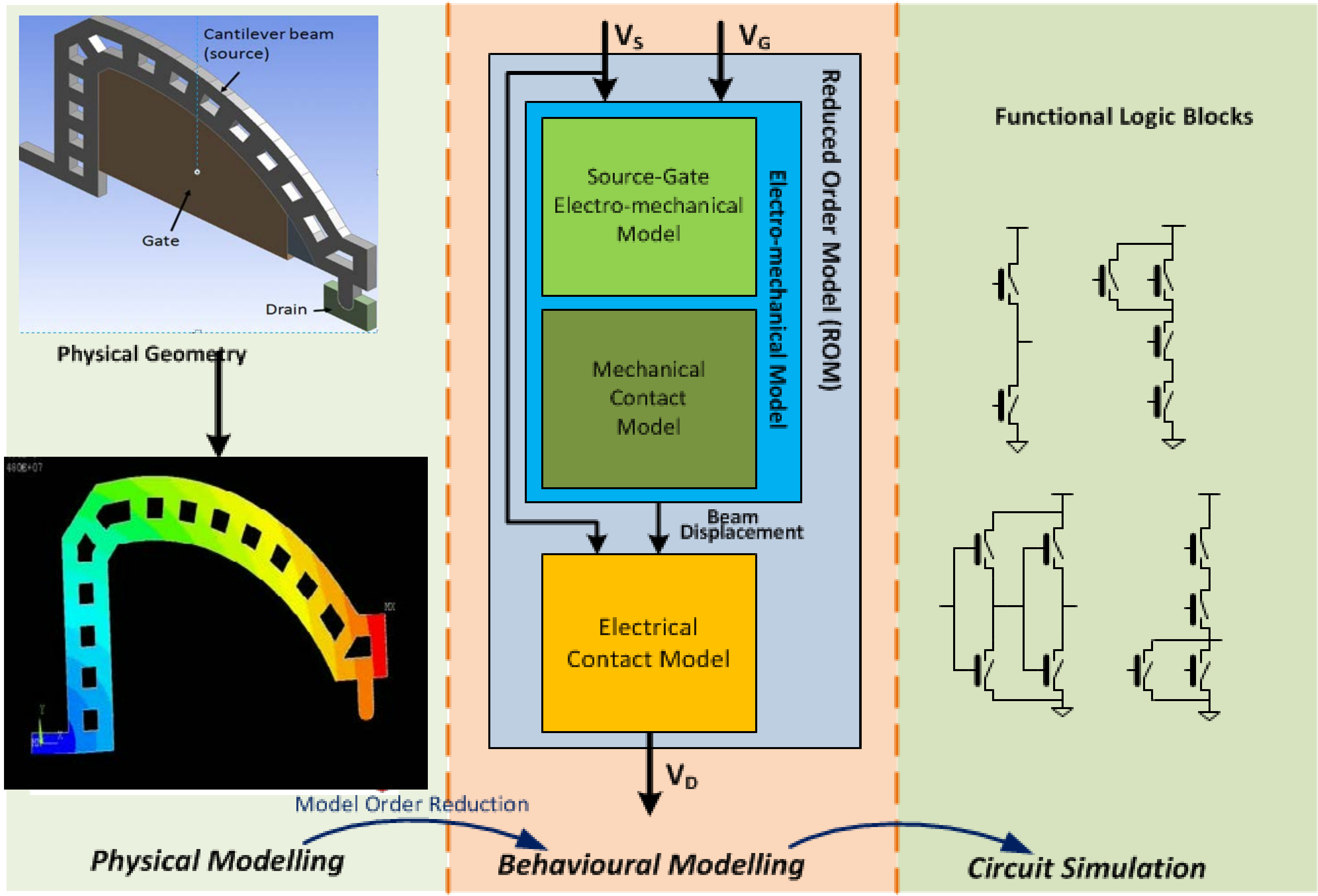
Long term integration plan
courtesy F. Niklaus et al. KTH, Sweden



[1] D. Grogg et al., "Curved cantilever design for a robust and scalable microelectromechanical switch," *Int. Conf. on Electron, Ion, and Photon Beam Technology and Nanofabrication*, Waikoloa, Hawaii, 2012.

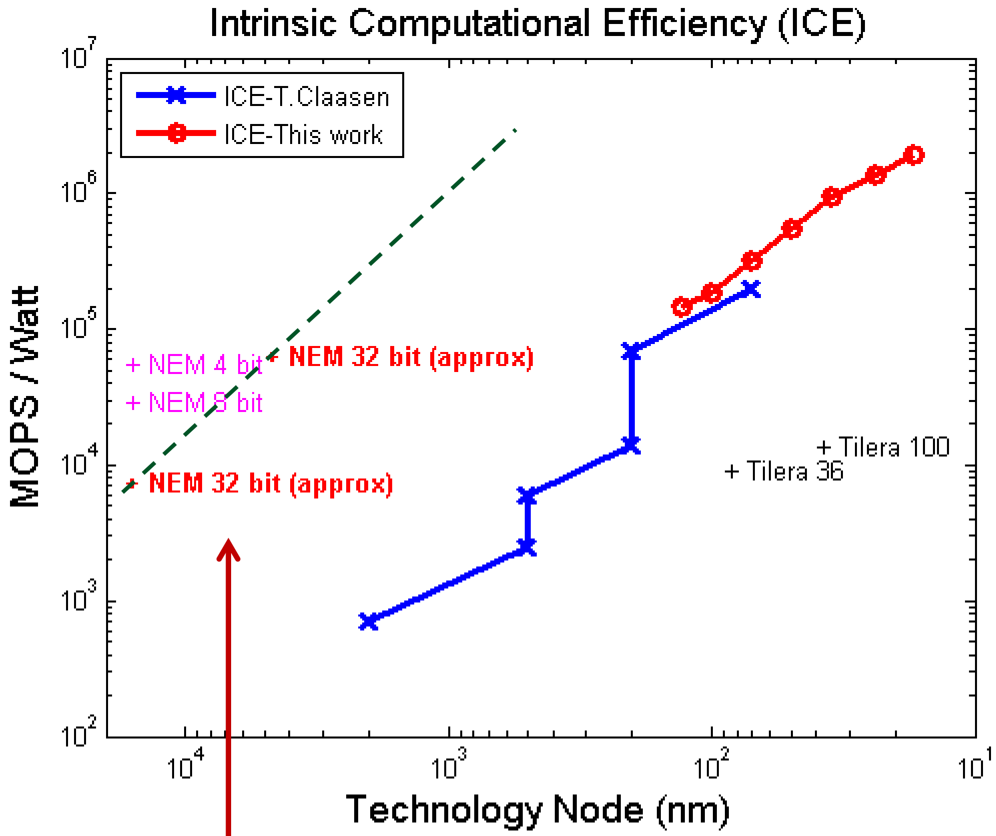
NEM Relay Modelling

❑ FEA to Reduced Order Model to Circuit Model



ICE with NEM Logic

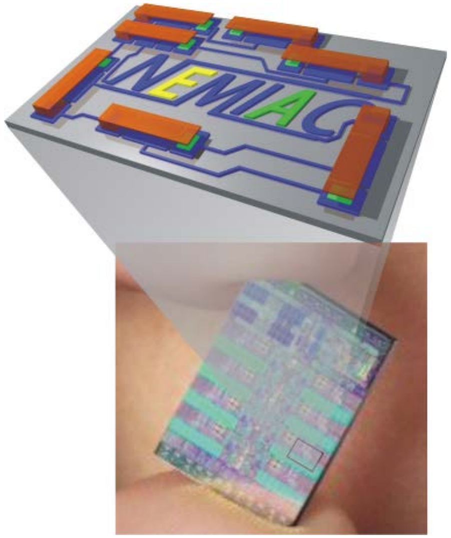
- **NEM "tech node" is much larger than CMOS**
  - Devices fabricated at 17 μm and 5 μm cantilever length
- **Miniaturisation increases speed and reduces energy**
  - No straightforward analogue to scaling for CMOS
- **Trajectory is promising**
  - Ultra-low power technology for low latency, low throughput applications



- Ripple-carry architecture and worst-case energy
- Device models at 17μm and 5μm silicon qualified
- 4 and 8- bit adder energy at 17 μm based on accurate circuit model
- 32-bit adders based on scaling

Acknowledgements

- NEMIAC project
  - EU FP7 Strep: Grant No. 288670



- ELITE Project
  - EU FP7 Strep: Grant No. 215030