Computational Limits in 3-D Integrated Systems

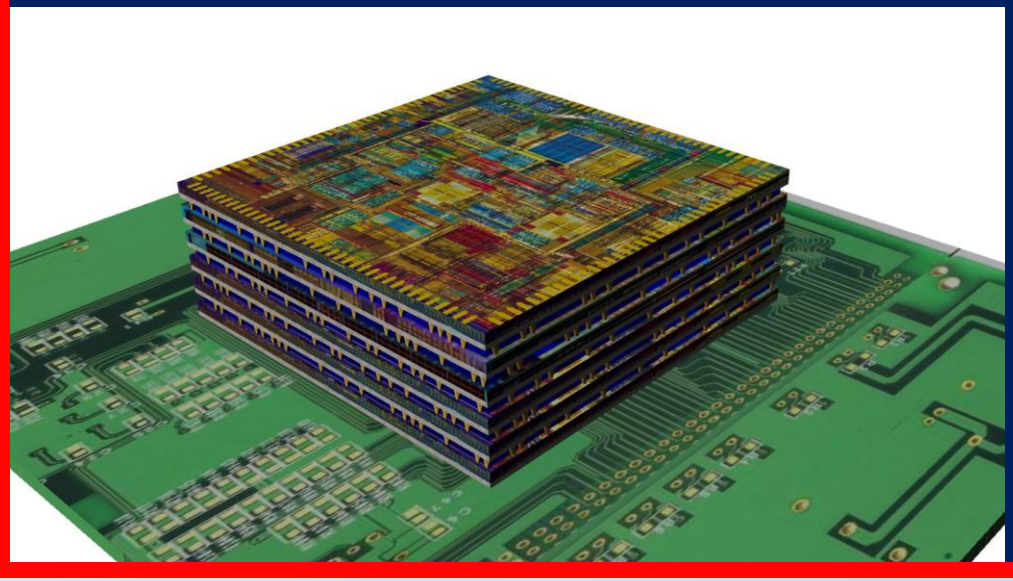
Axel Jantsch

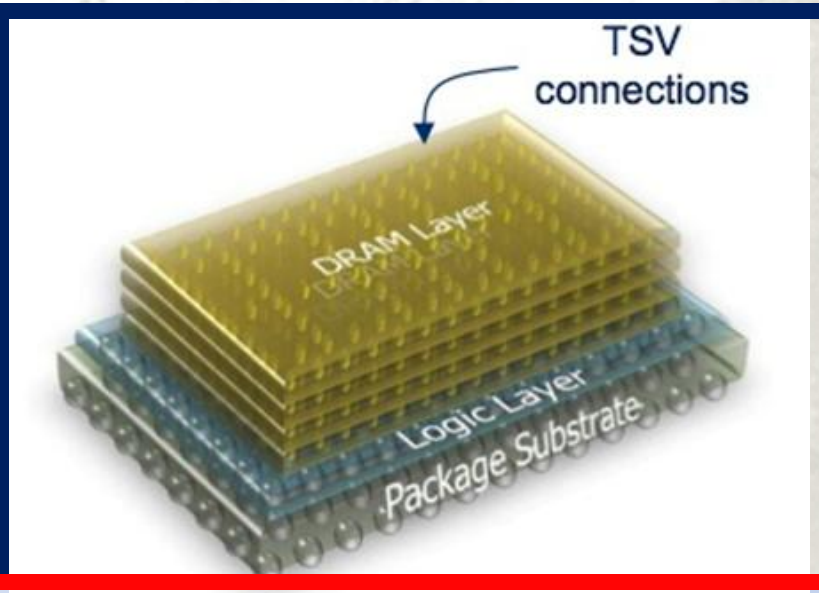
KTH Royal Institute of Technology - Stockholm, Sweden

Dinesh Pamunuwa, Matt Grange, Roshan Weerasekera*

University of Lancaster - Lancaster, UK

* now at Agency for Science, Technology and Research (A*STAR), Singapore



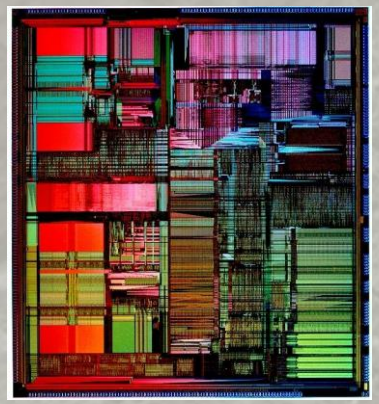


Presentation Outline

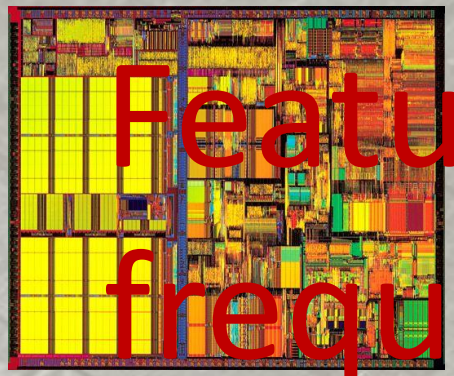
- **Introduction**
  - Overview
- **Modelling computational efficiency**
  - Methodology
  - Architectural abstraction parameters
- **Underlying physical Models**
  - Logic, Interconnect, Memory
  - Thermal behaviour
- **Applications of the Models**
  - Modelling implemented processors
  - Effect of system architecture on computational efficiency
- **Conclusions**

The Evolution of Single Die Processors

LANCASTER UNIVERSITY



Pentium I [800 nm]

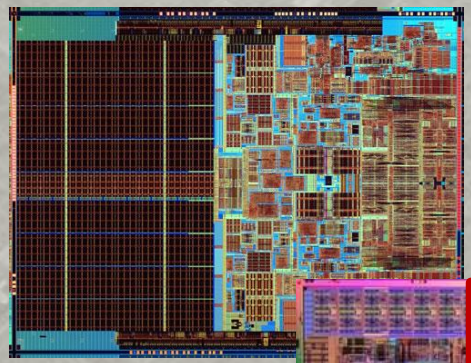


Pentium II [350 nm]

Pentium III [180 nm]

Pentium IV [130 nm]

Feature size and frequency SCALING!



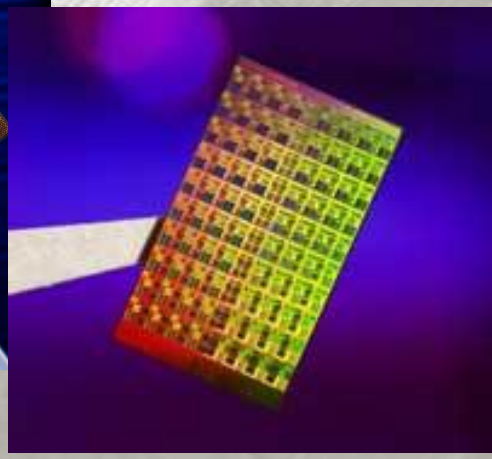
Intel core2duo [2 core]

Intel i7 [4 core]

Xeon [8 core]

PARALLELISM!

Intel SCC [48 core]



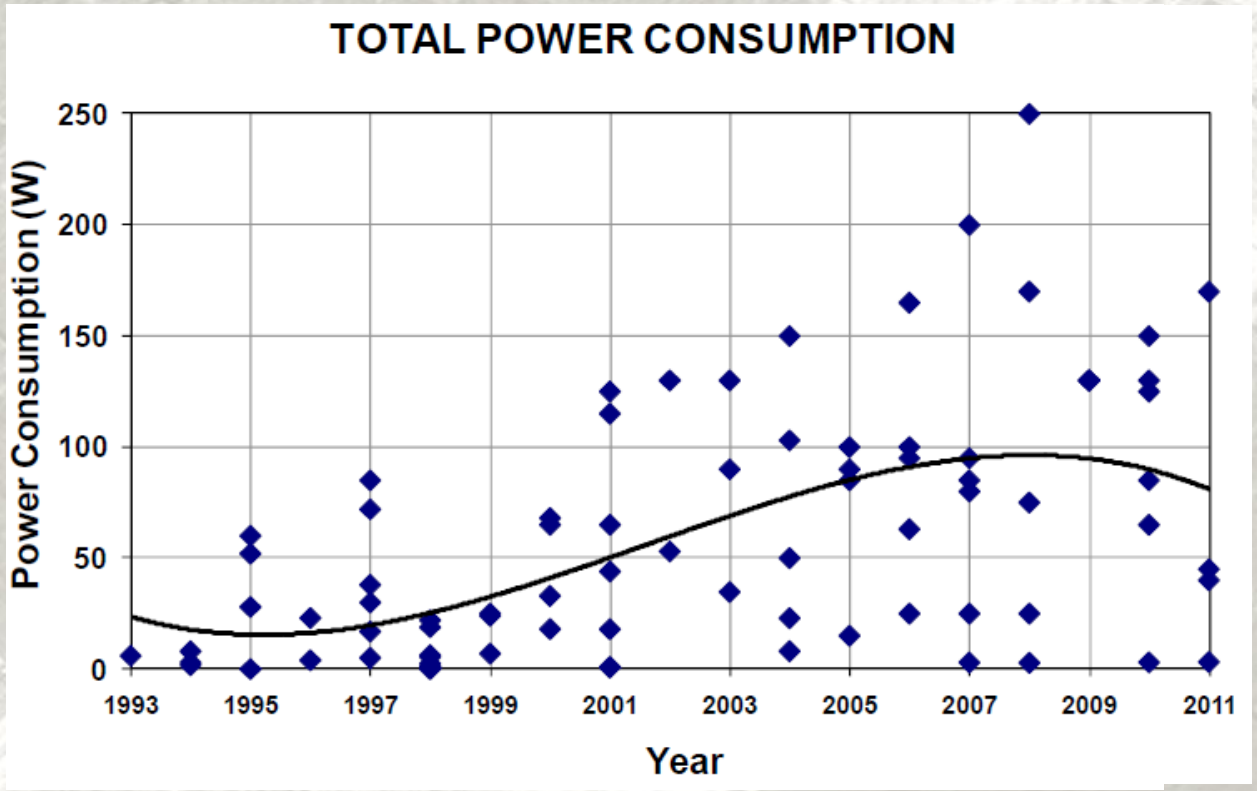
Intel Experimental [80 core]

(micrograph images: Intel)

Technological Advancement

LANCASTER UNIVERSITY

- In the past, feature size, V_{dd}, and clock frequency scaling has allowed processor performance to increase favourably
  - The Intel 4004 in 1972 had 2,300 transistors clocked at 108 KHz.
  - The 2011 Intel Itanium server processor crams 3.1 billion transistors clocked 2 GHz.
  - [illegible] benefits



(Source: ISSCC Trends 2011)

Future Direction of High Performance ICs

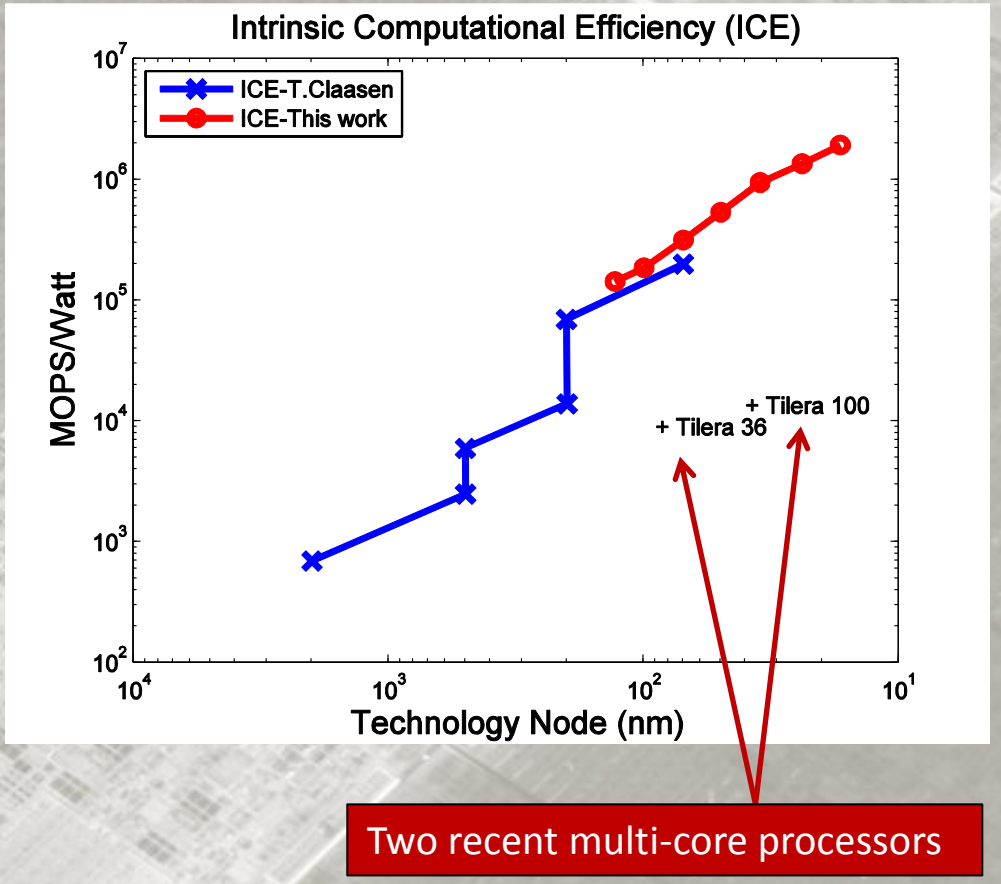
- **Advantages of 3-D**
  - TSVs have much better electrical characteristics: faster, less energy
  - Allows heterogenous integration: DRAM in same package
  - Average interconnection length reduces: less energy
- **Disadvantages of 3-D**
  - Thermal dissipation harder: lower overall power
  - Cost: potentially higher

Presentation Outline

- Introduction
  - Overview
- **Modelling computational efficiency**
  - Methodology
  - Architectural abstraction parameters
- Underlying physical Models
  - Logic, Interconnect, Memory
  - Thermal behaviour
- Applications of the Models
  - Modelling implemented processors
  - Effect of system architecture on computational efficiency
- Conclusions

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.



Expanding the *ICE* to the *ECE*

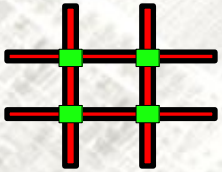
- The *ICE* gives the *maximum upper bound* on efficiency, but cannot account for realistic systems because it only considers the computational unit.

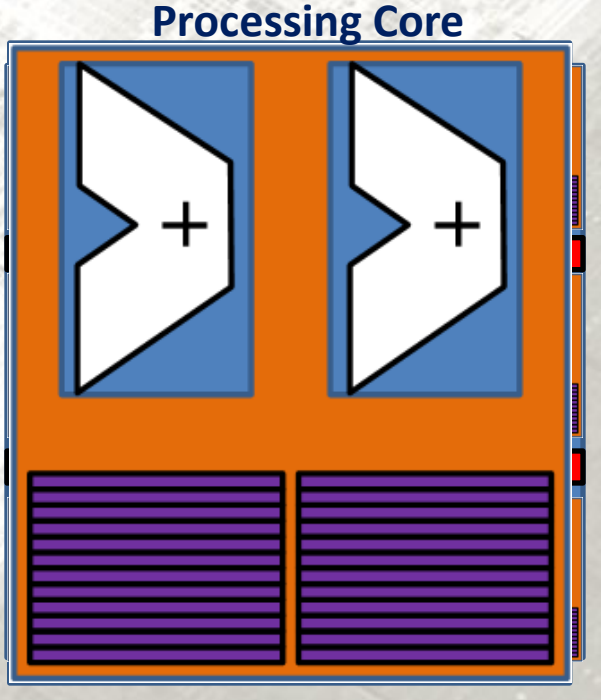
- We build upon the *ICE* by modelling the three fundamental operations of any processing unit:

- The computational operation

- The memory

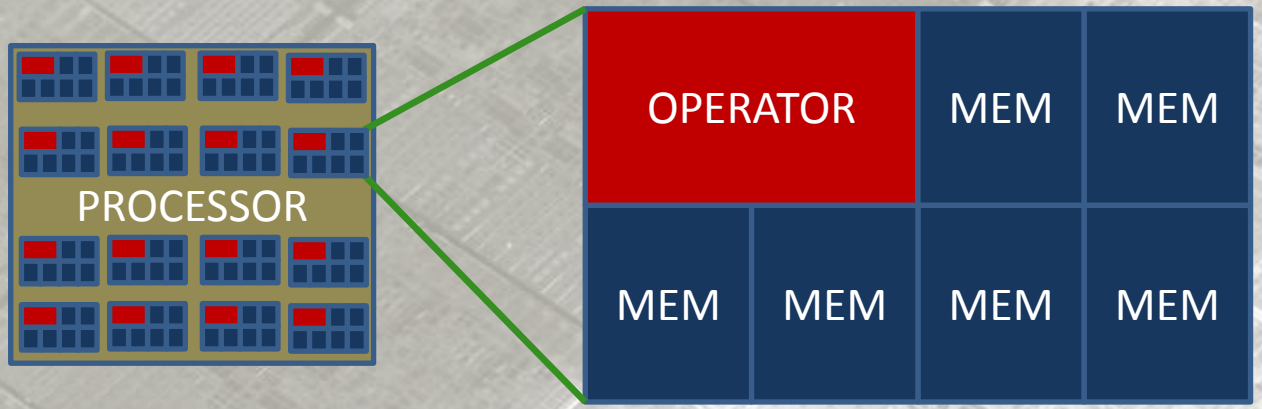
- The interconnect





Temporal and Spatial Organization of Memory

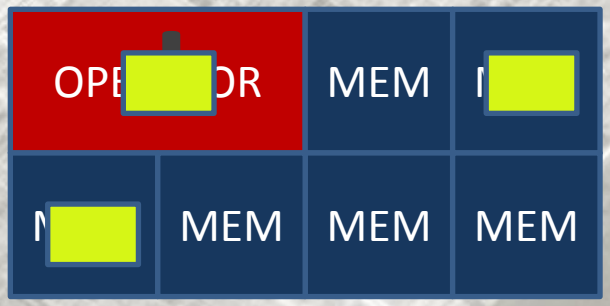
- μ_s
  - Gives us the amount of memory per operator. Think of it as the amount of on-chip cache available



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

2 reads, 1 write

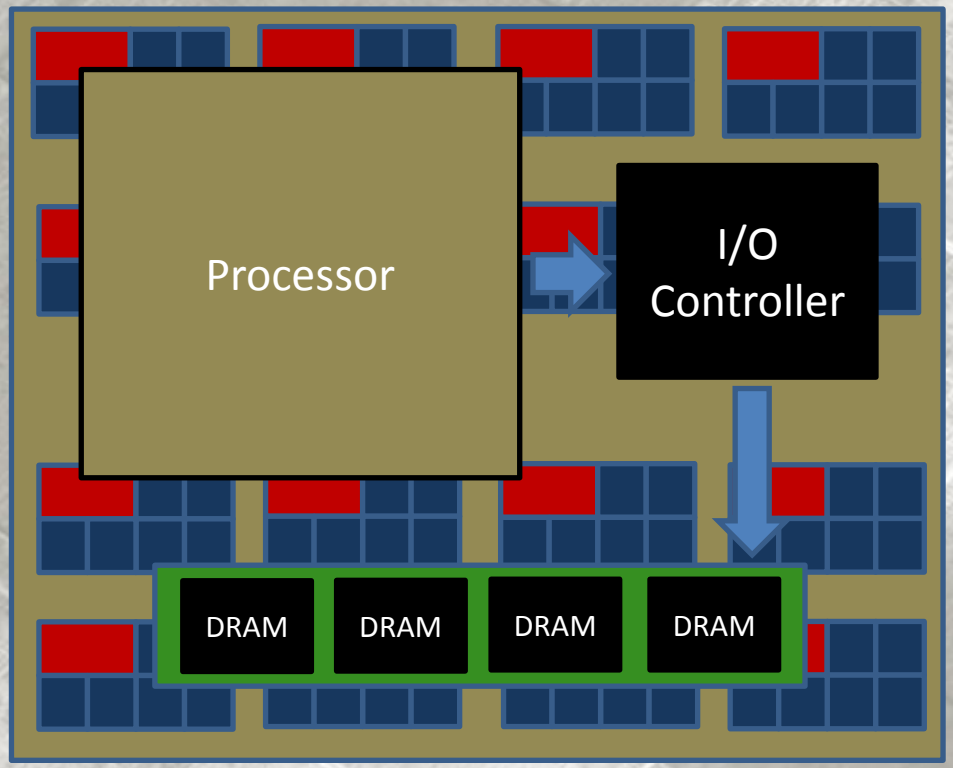
$\mu_T = 3$



On- or Off-chip Memory?

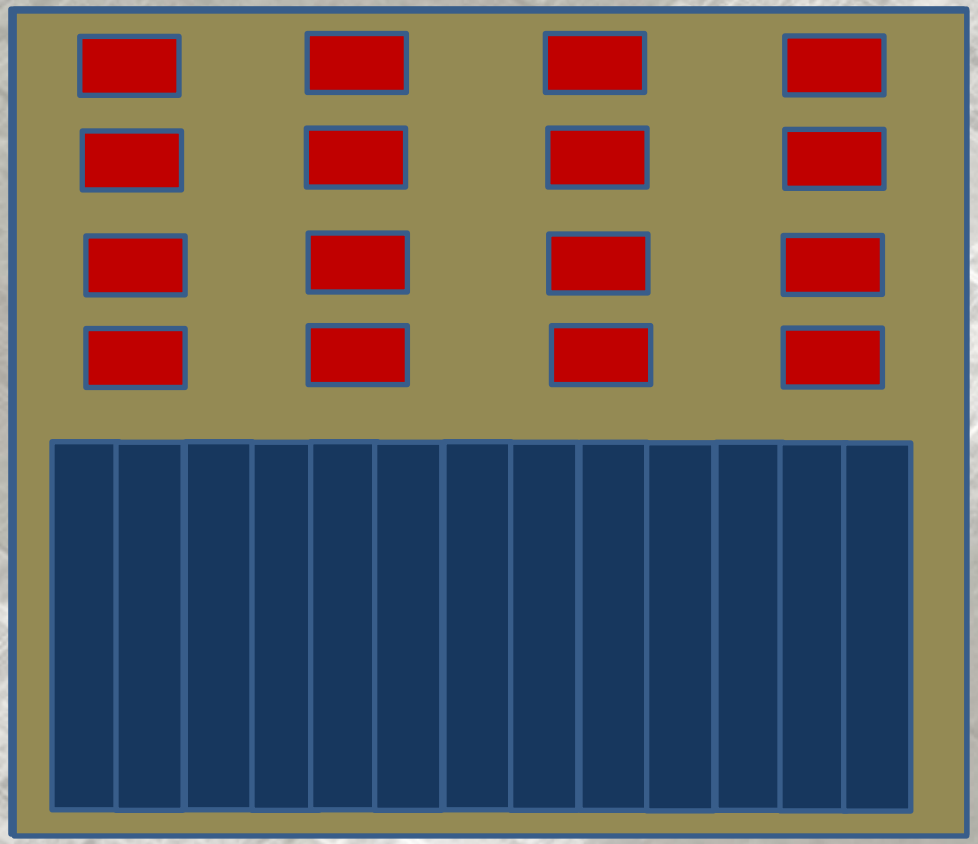
- ω (0-1)
  - Gives the ratio of on- to off-chip memory in the system. Off-chip memory requires exiting the die with I/O drivers and external chips.

0: all-off-chip



Memory Distribution Factor

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



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

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

Ratio of on-chip : off-chip memory

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

Amount of computation possible within the envelope of 1 Joule

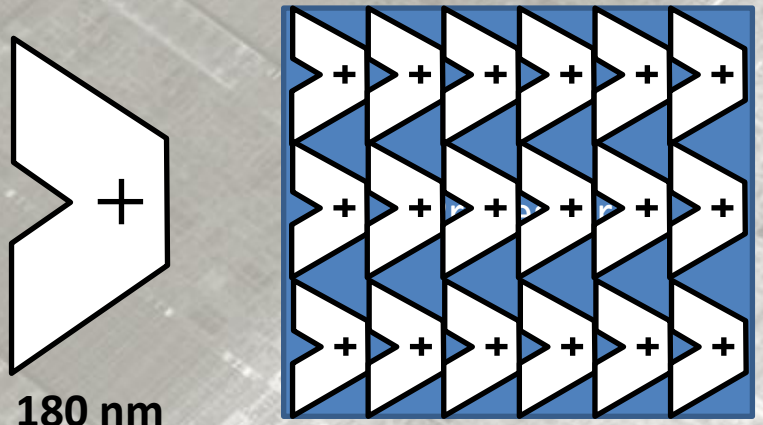
Presentation Outline

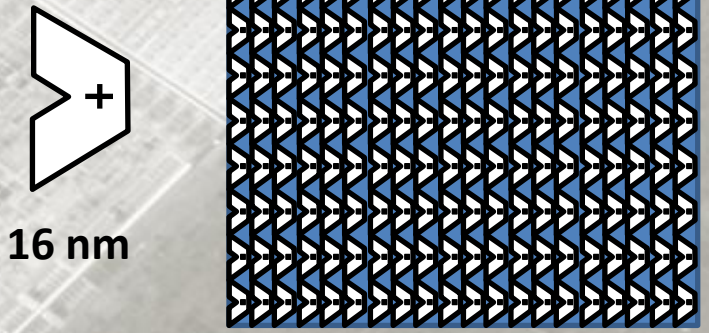
- Introduction
  - Overview
- Modelling computational efficiency
  - Methodology
  - Architectural abstraction parameters
- Underlying physical Models
  - Logic, Interconnect, Memory
  - Thermal behaviour
- Applications of the Models
  - Modelling implemented processors
  - Effect of system architecture on computational efficiency
- Conclusions and Future Work

Computation: 32-bit Adders

- **The fundamental computational operation we use is a 32-bit add operation.**
  - **We use published data for energy and area for a 32-bit adder implemented in one technology node and scale the dynamic energy and area according the following:**

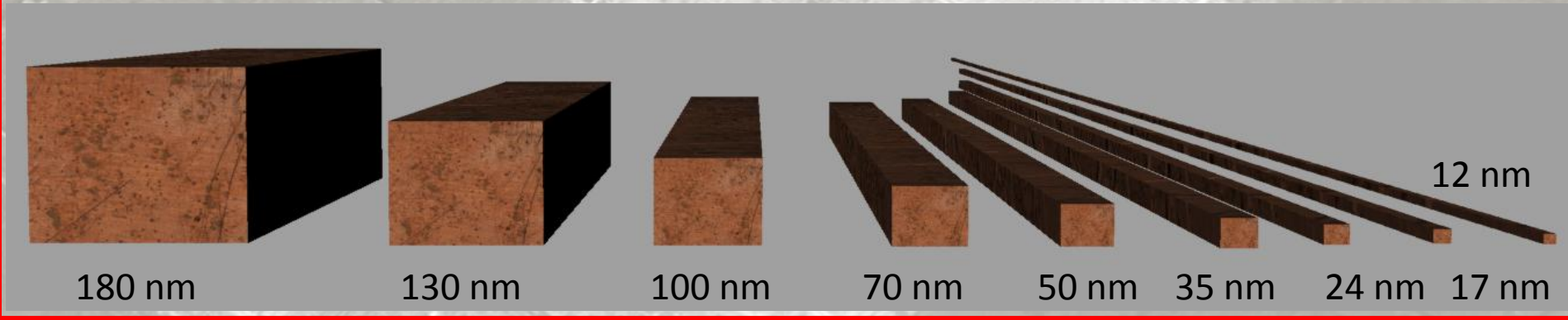
$$Energy_{new} = Energy_{130nm} \times \left(\frac{Feature_{size_{new}} \times Vdd_{new}^2}{Feature_{size_{130nm}} \times Vdd_{130nm}^2} \right).$$





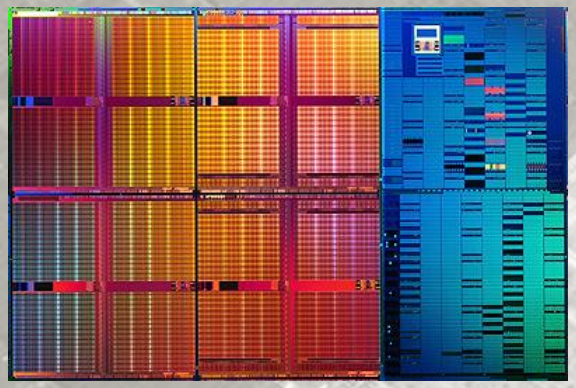
Interconnect: Scaling and TSVs

- **Feature size scales by roughly 0.7 each generation, but global wires do not scale as aggressively.**
  - **We have tabulated RC parasitics for global wires and drivers across technology generations.**
  - **The energy-per-bit of sending an n-bit word, any length, has been calculated including driver, receiver and repeater energy.**
  - **TSV parasitics are extracted from a field solver for lengths of 50 µm and radii of 1, 2, 4, 8, and 10 µm**



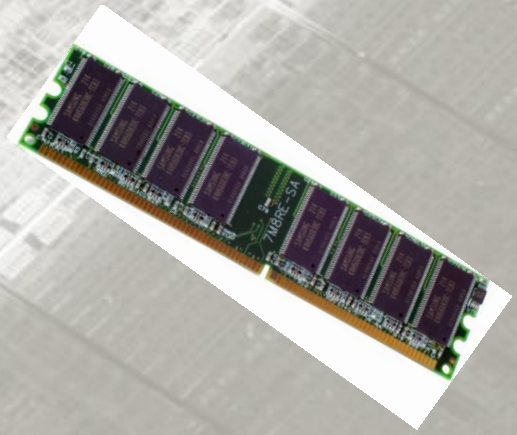
Memory: DRAM and SRAM

- **To maximize the benefits of 3-D ICs, we consider DRAM in the stack, which is much higher density than typical SRAM used in most 2-D ICs.**
  - **We have scaled both SRAM and DRAM (DDR2-4) in terms of energy and latency in a similar manner to the logic.**
  - **We also have modelled off-chip DRAM using the Micron System Power Calculator.**



On-chip SRAM
(image: Intel)

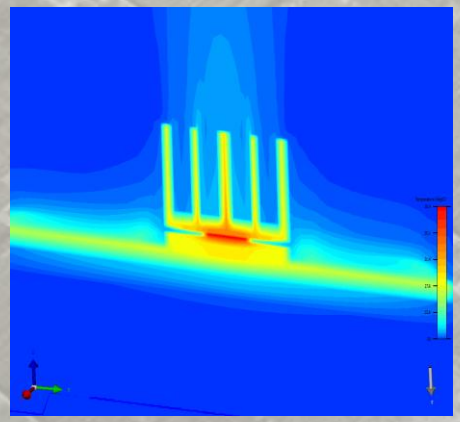
On-chip stacked DRAM
(image: samsung)

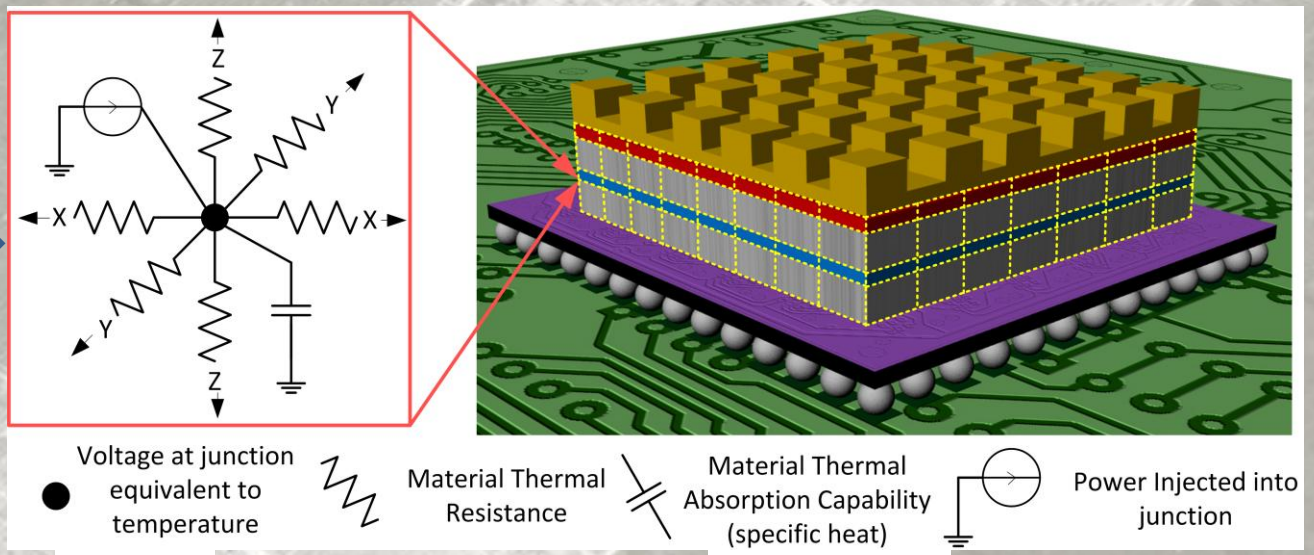


Off-chip DRAM
(image: micron)

Thermal Behaviour

- Simulations have been conducted using a Computational Fluid Dynamic (CFD) solver (FloTHERM).
- The thermal behavior and limitations of 2-D and 3-D packages have been extracted from simulations.
- Compact thermal models have been developed as part of the toolset to quickly predict the thermal behavior.





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

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

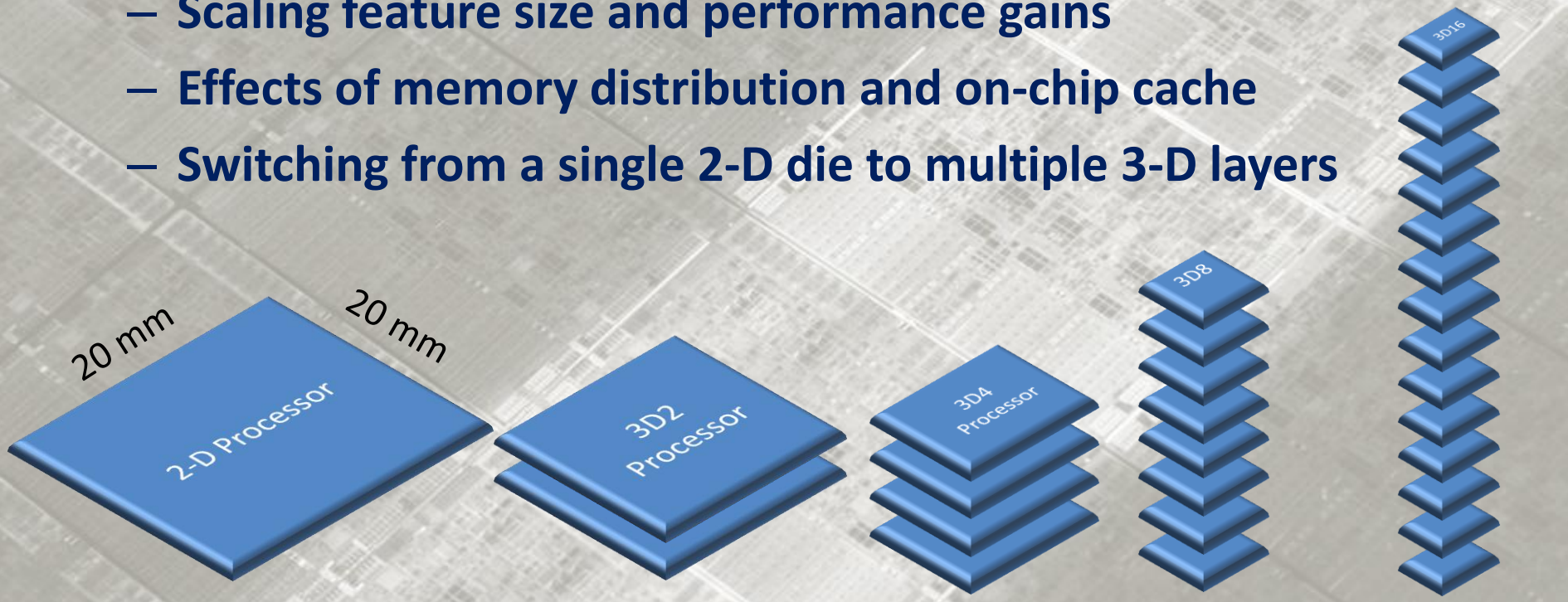
$$\theta_{j-a} = \frac{T_j - T_a}{q}$$

Presentation Outline

- Introduction
  - Overview
- Modelling computational efficiency
  - Methodology
  - Architectural abstraction parameters
- Underlying physical Models
  - Logic, Interconnect, Memory
  - Thermal behaviour
- Applications of the Models
  - Modelling implemented processors
  - Effect of system architecture on computational efficiency
- Conclusions and Future Work

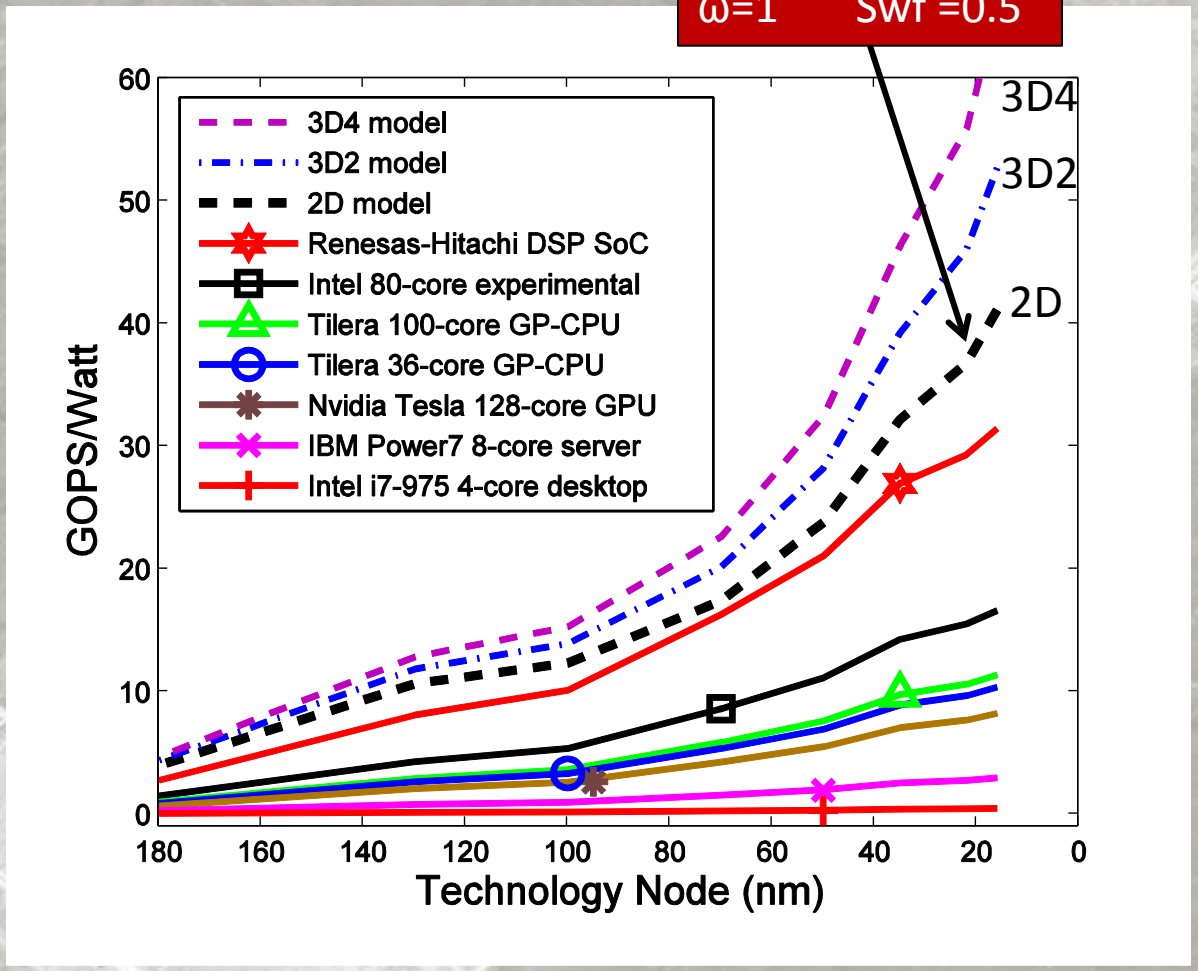
System Comparison

- **We choose a few examples to demonstrate the differences between 2-D and 3-D ICs to examine the impact of several design choices:**
  - **Scaling feature size and performance gains**
  - **Effects of memory distribution and on-chip cache**
  - **Switching from a single 2-D die to multiple 3-D layers**



ECE vs. Technology and Architecture

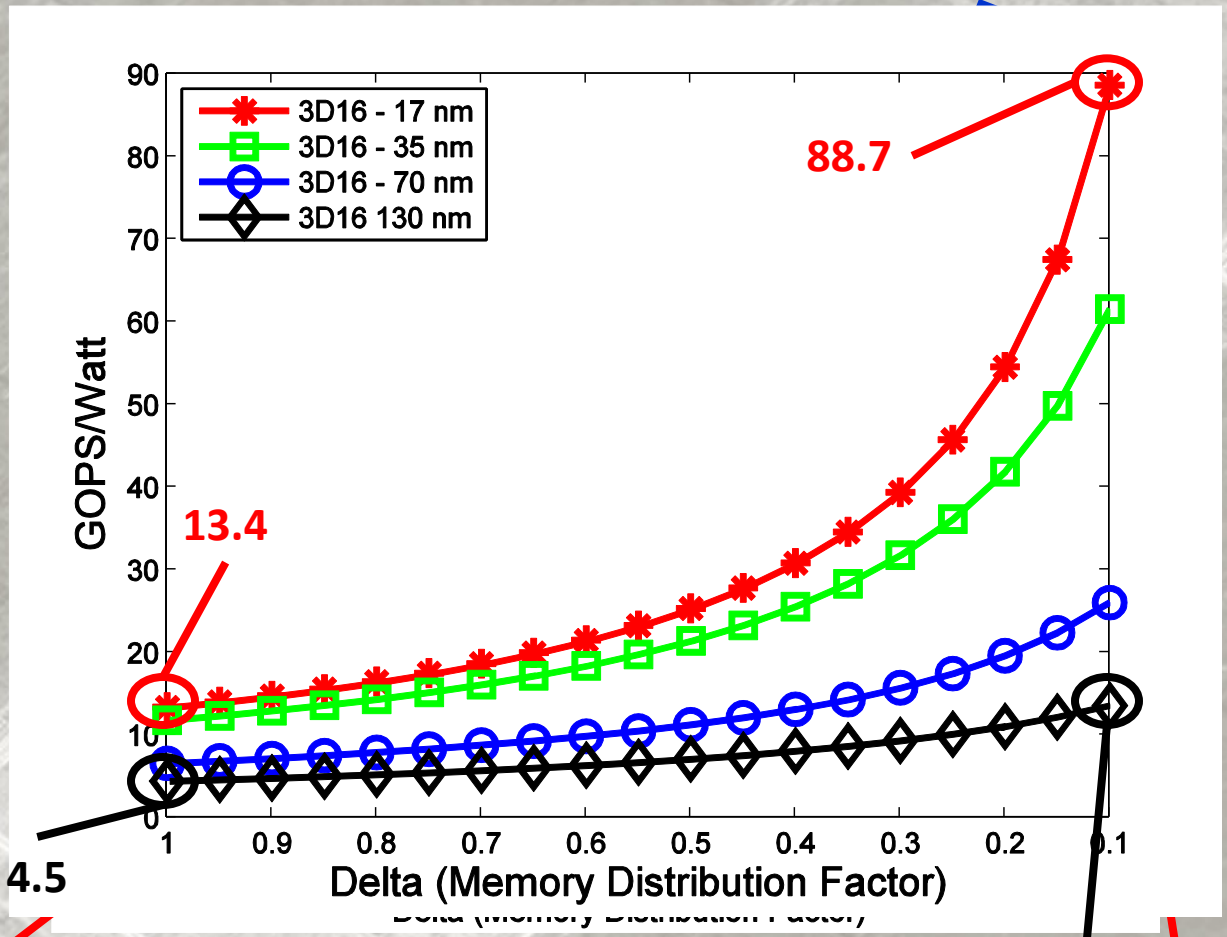
The model parameters can be altered to represent virtually any system, where the difference between the maximum and the actual implementation is dictated by the efficiency of the control circuitry.



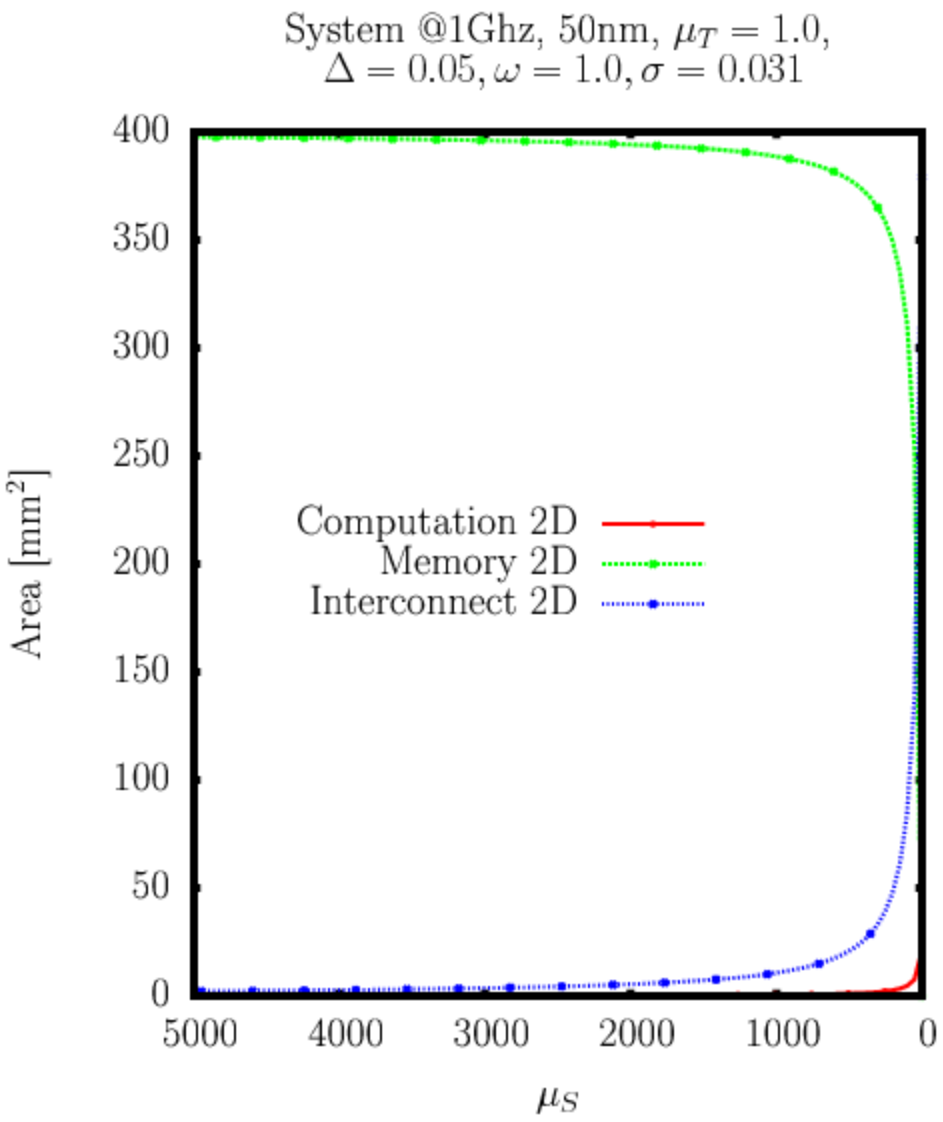
Memory Distribution Factor

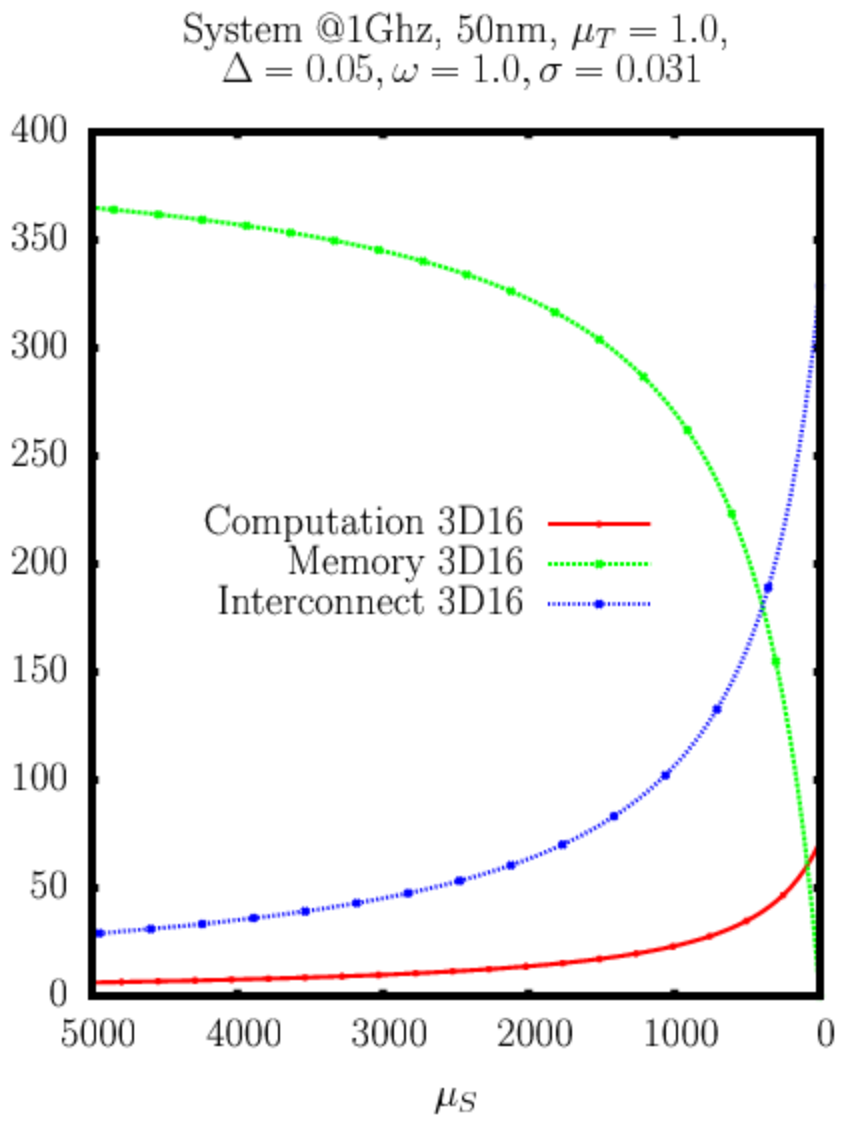
The advantage of 3-D systems reduces as the on-chip memory is moved closer to the computational units.

The advantage of distributing the memory grows at deep sub-micron dimensions

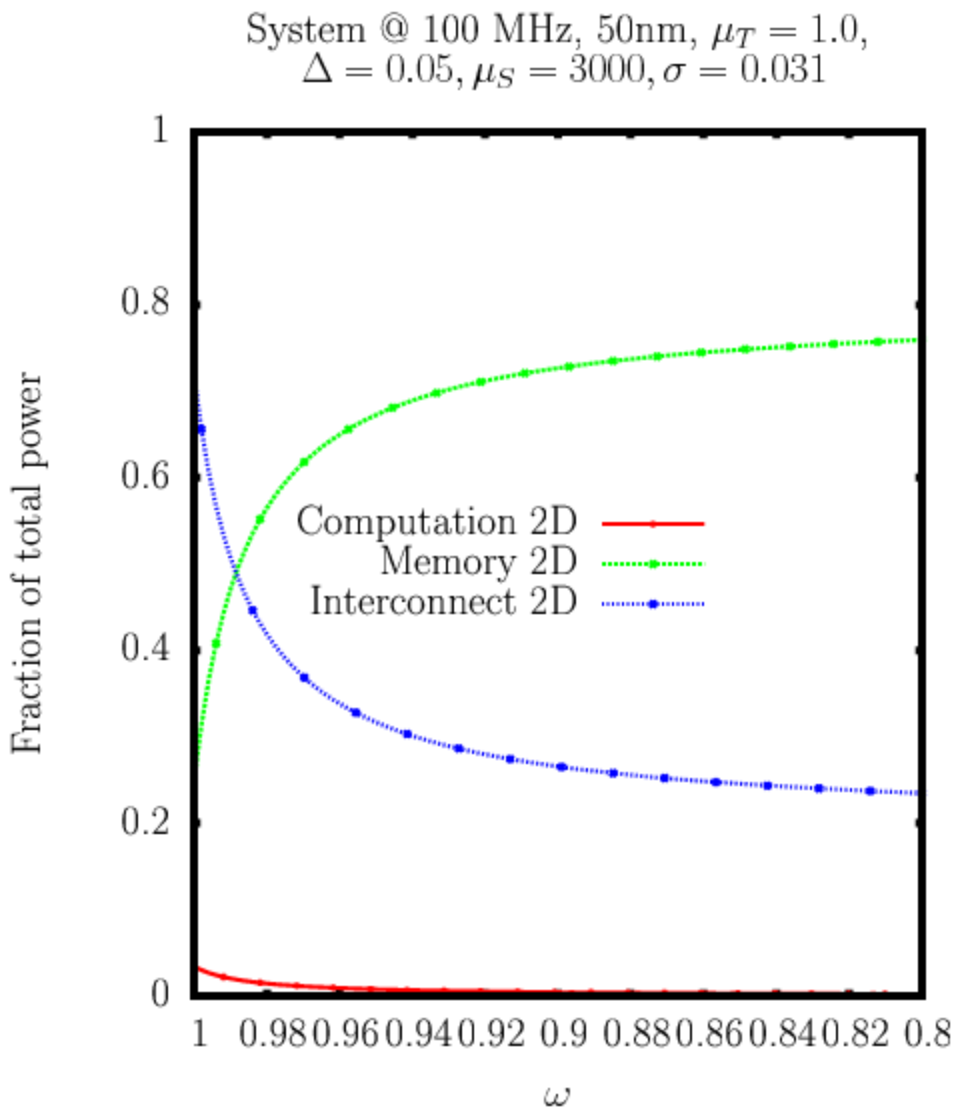


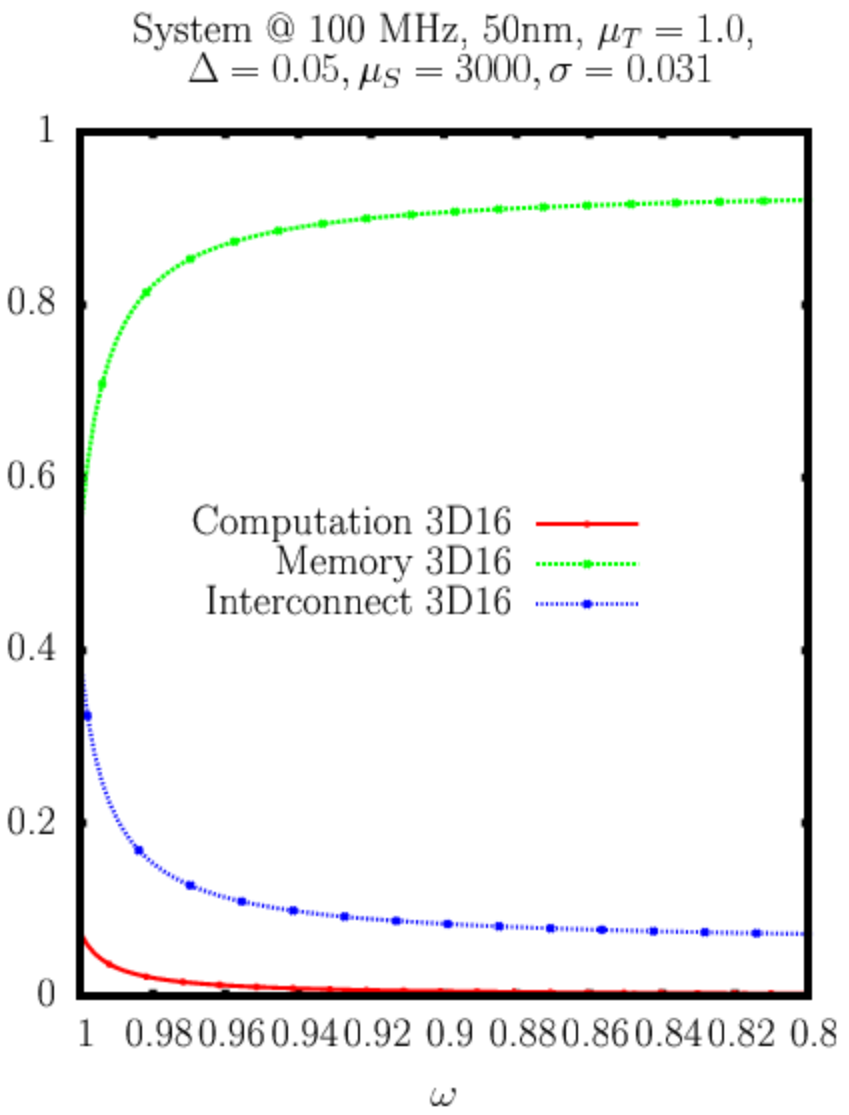
Area Distribution - 400mm^2



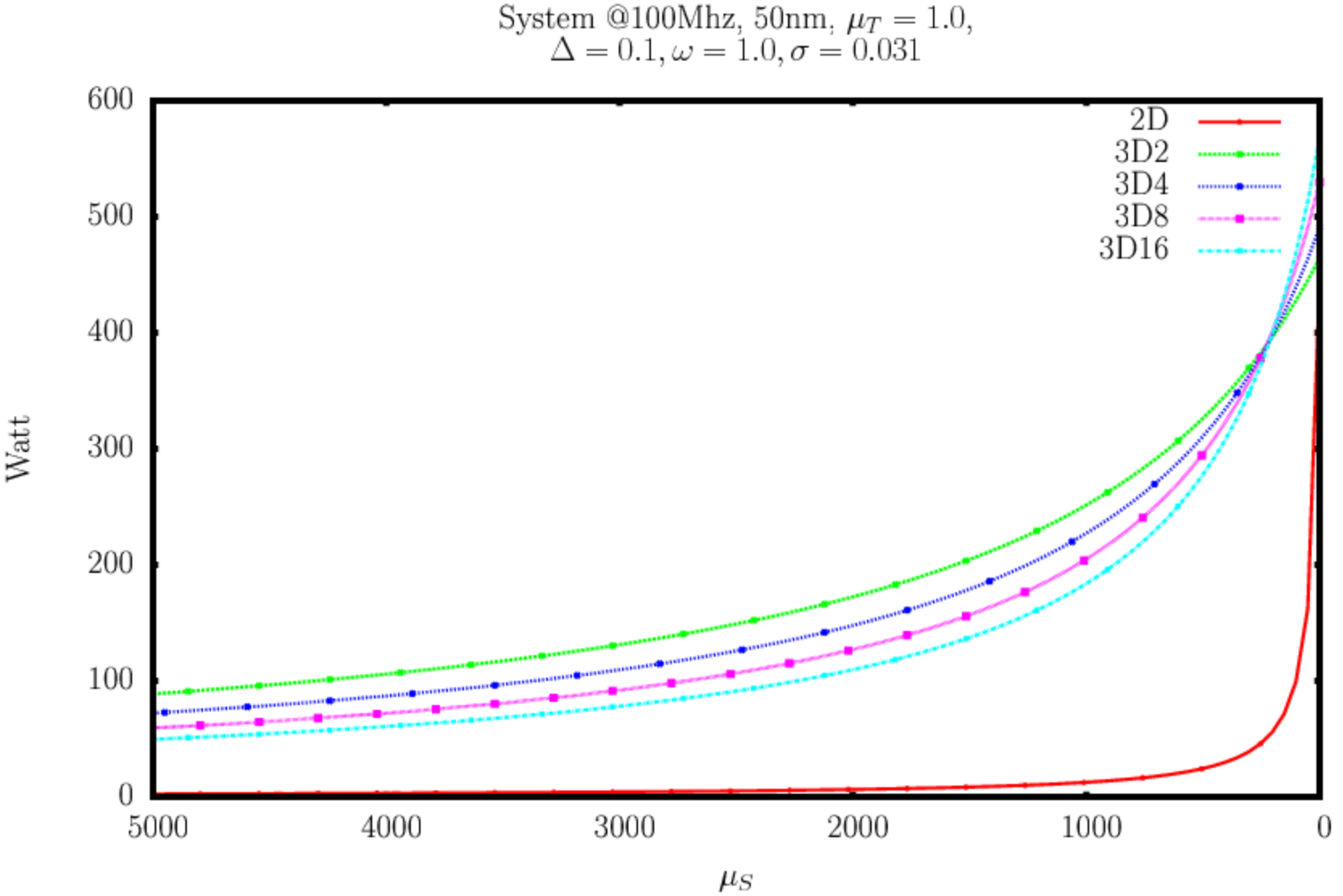


Power Distribution - 400mm^2

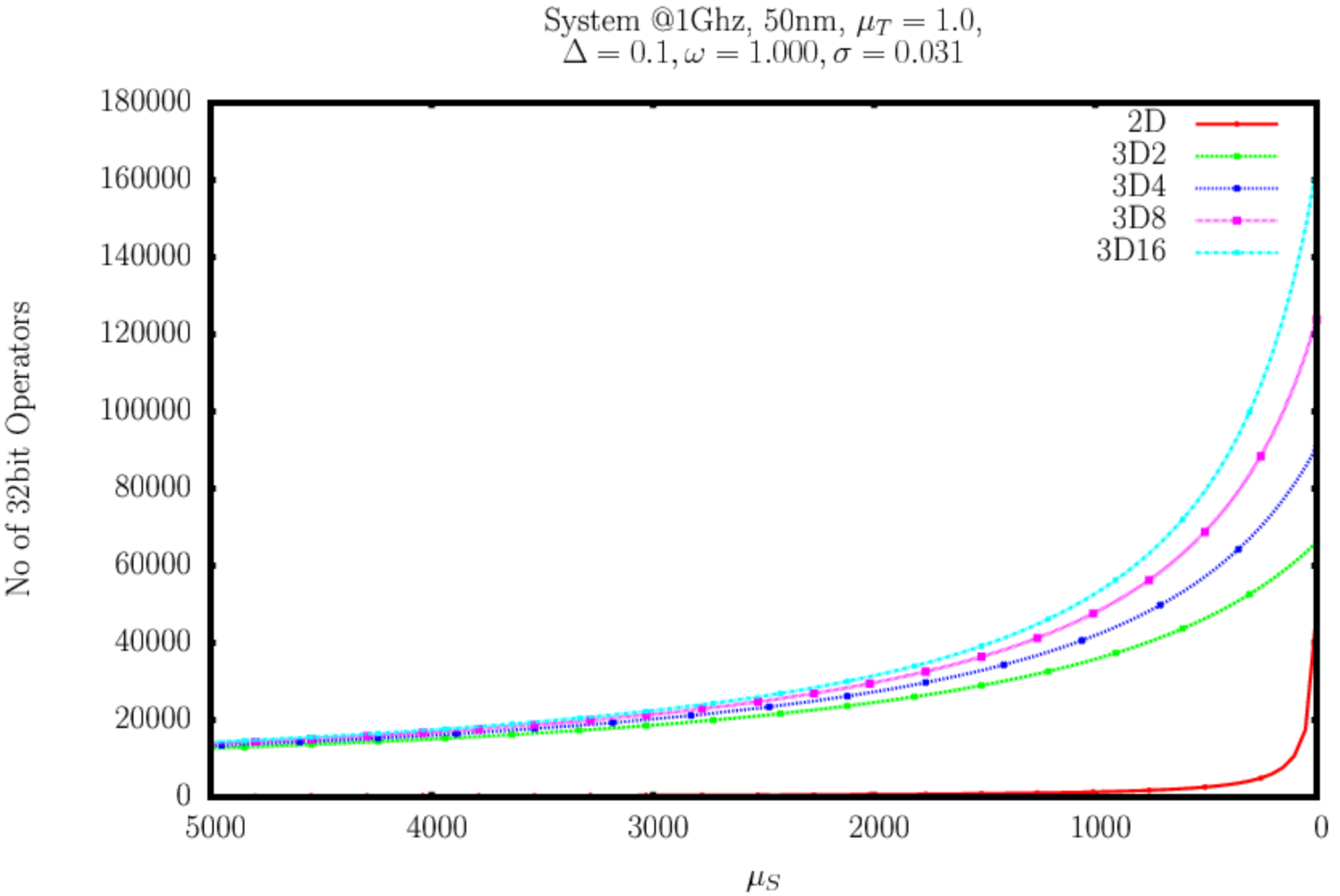




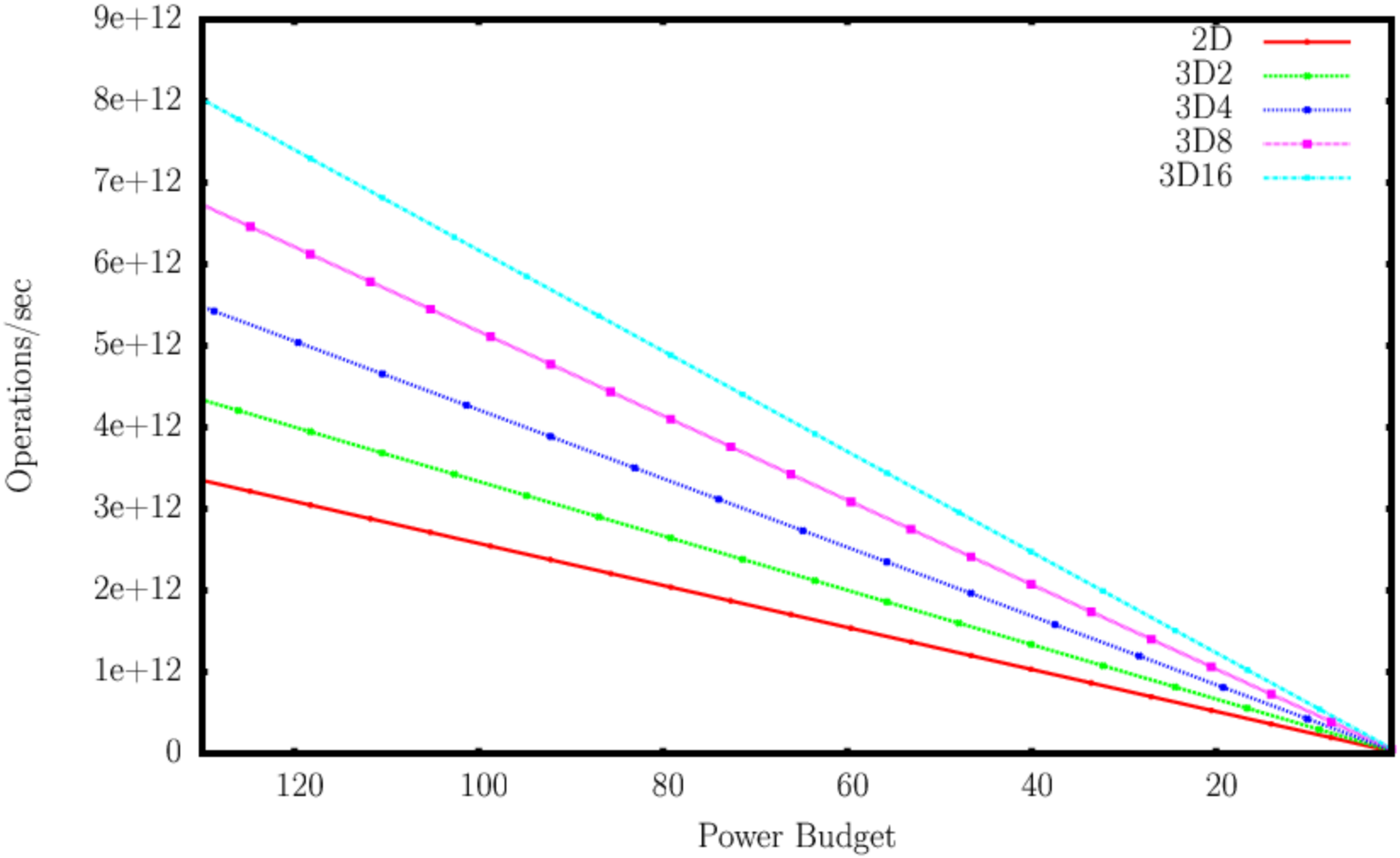
Power Consumption in 400mm²



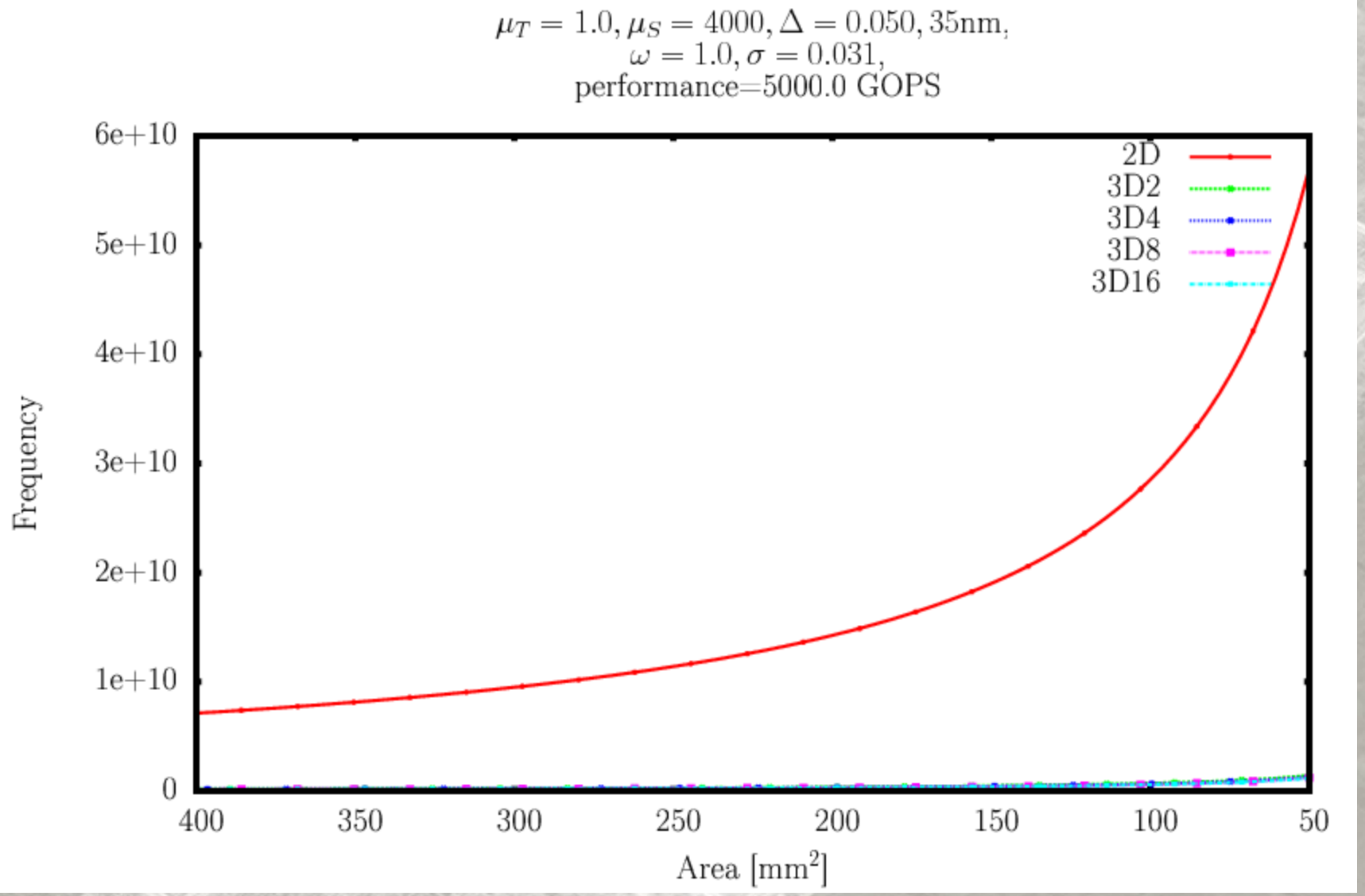
No of operators in 400mm²



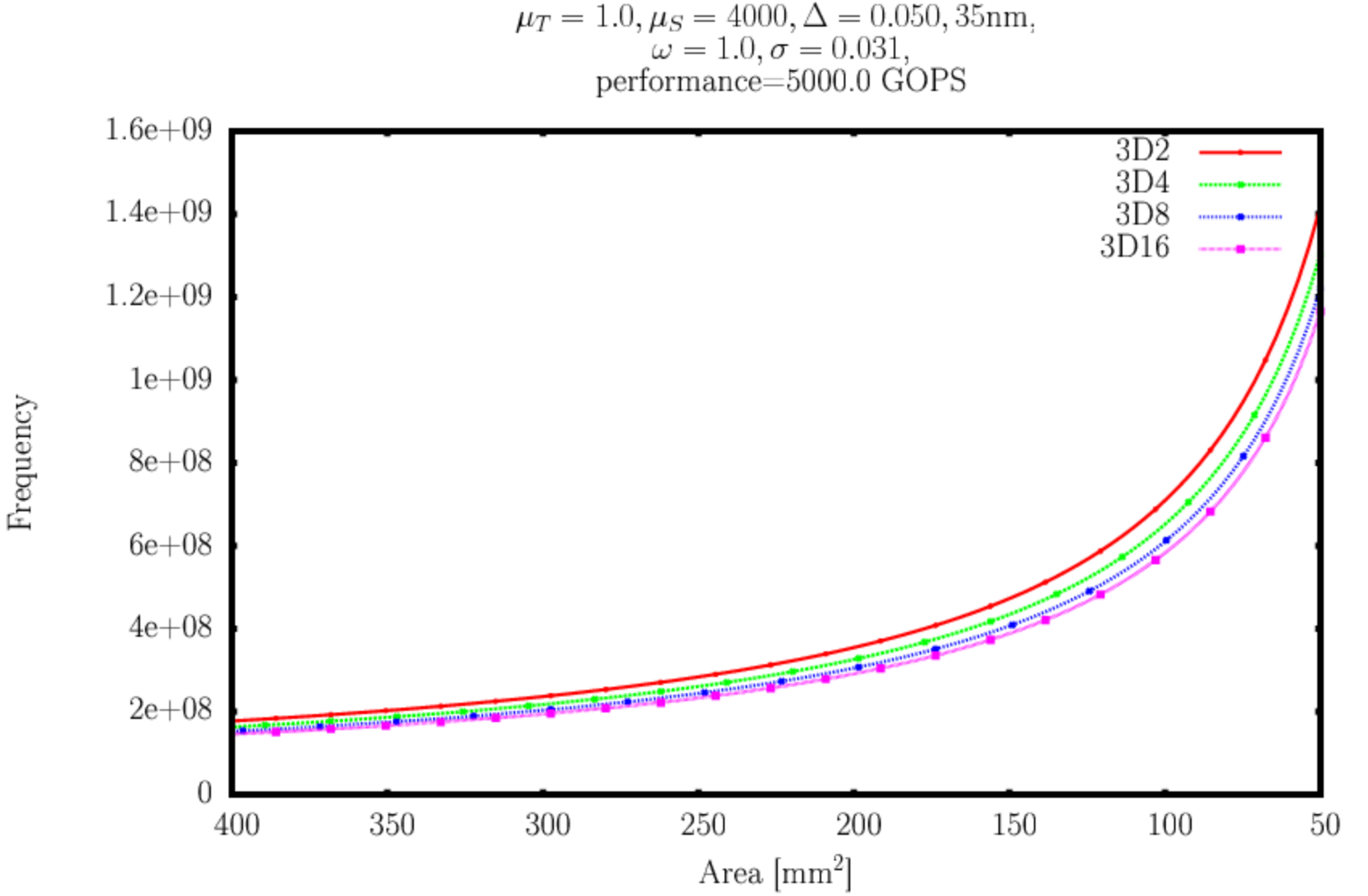
Performance under Power Constraint



Frequency to Obtain a Given Performance



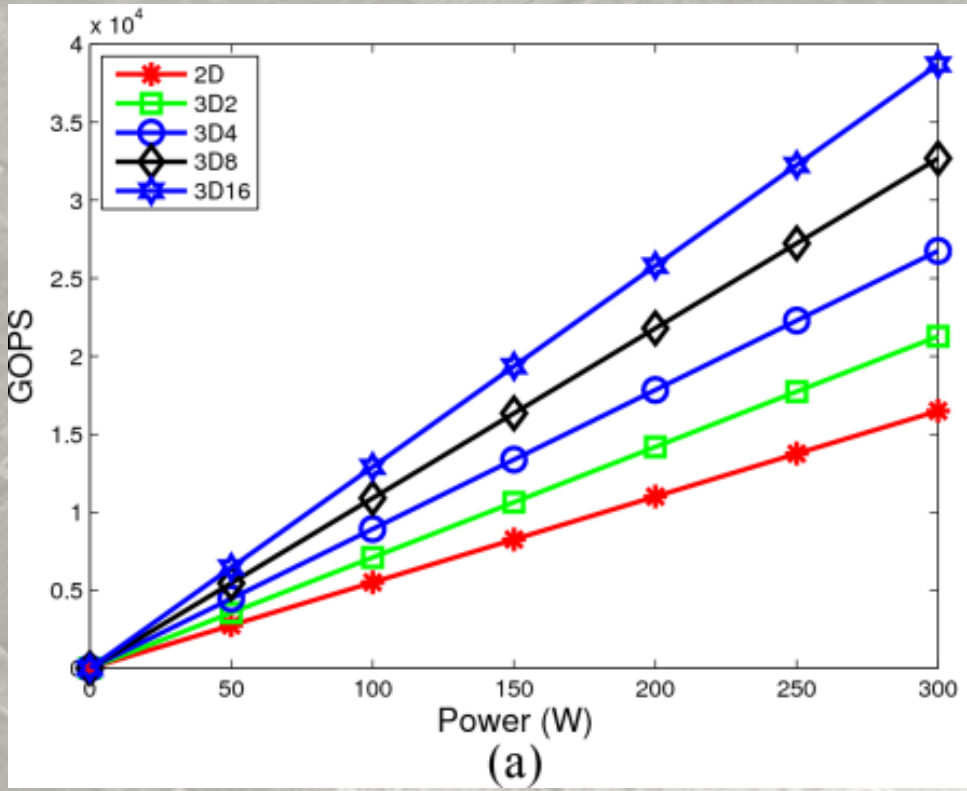
Frequency to Obtain a Given Performance

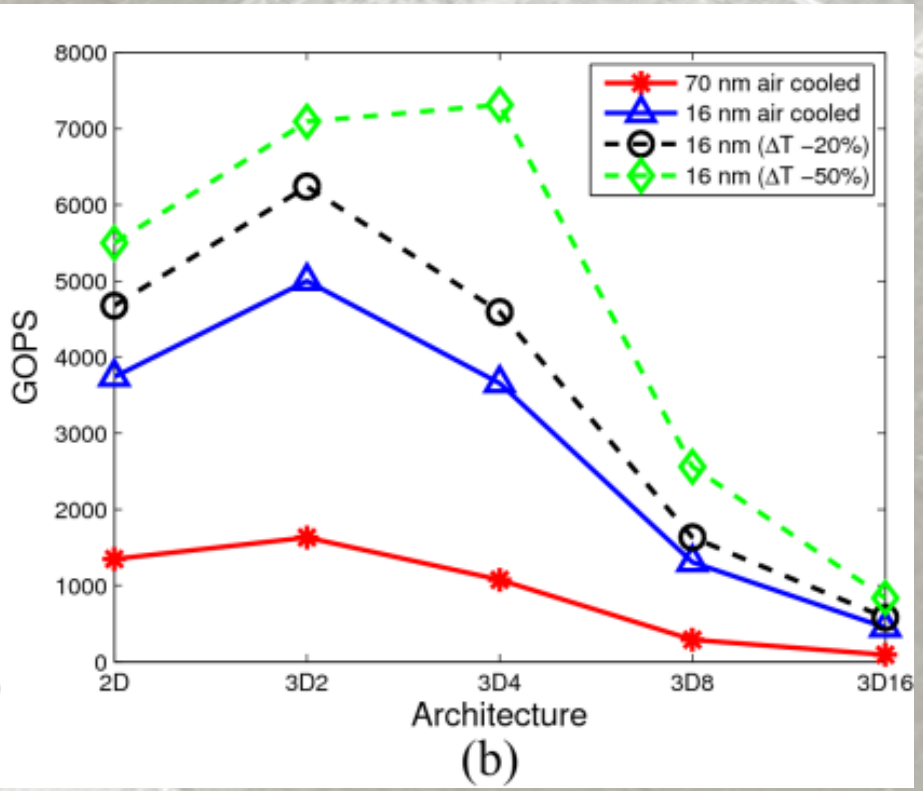


Power and Thermal Limitations

Performance under power constraint

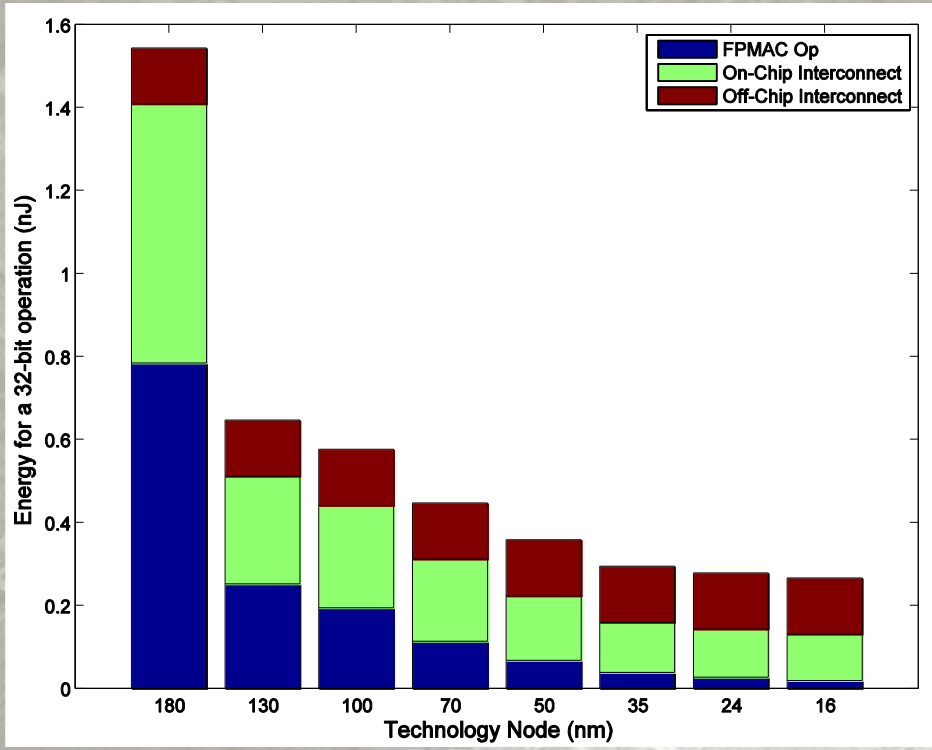
Performance under temperature constraint



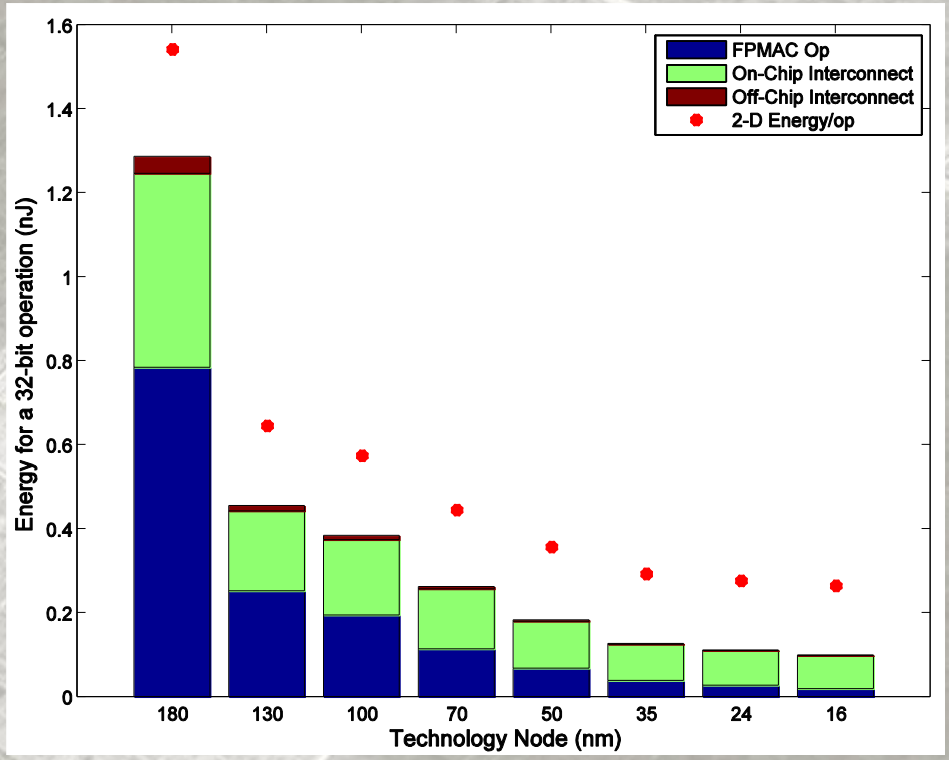


Scaling 2-D versus 3-D DRAM

2-D with off-chip DRAM



2-layer 3-D with in-stack DRAM

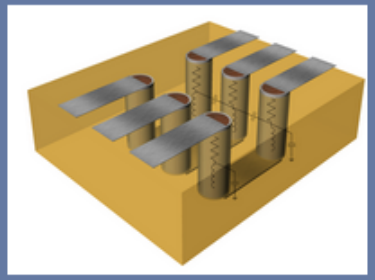


3-D IC Modelling Tools

Tools for Design Space Exploration of 3-D Integrated Circuits

http://3d-performance.lancs.ac.uk/

Parastic Parameter Extraction of TSVs



Parasitic parameter extraction of TSV structures is a critical first step in estimating delay, signal integrity (SI) and power integrity (PI) of circuits in the design and verification of 3-D ICs. We have developed a set of compact models for resistance, (self and mutual) capacitance and (self and mutual) inductance of TSVs based on user-defined geometric configurations as well as different substrate types within an appropriate equivalent circuit for chip planning and design space exploration. This is available as a web-based tool here.

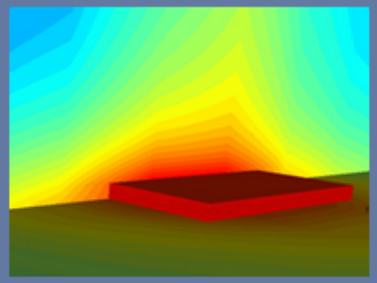
Link to tool: Parasitic Extraction

Documentation

Research paper describing models

Help file

Thermal Modelling of 3-D ICs

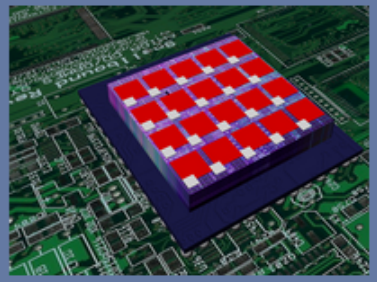


Heat dissipation and thermal management within a die stack is a critical issue in 3-D IC design. Of especial concern is the possibility of thermal runaway, where increased heat leads to increased leakage and further heat generation, which becomes a positive feedback cycle potentially leading to catastrophic failure. We have implemented a 2-D heat model to predict the temperature within the die stack depending on user-specified geometry and thermal interposer materials as well as heat sinks. This is available as a web-based tool here.

Link to tool: Thermal Estimation

Documentation

Performance Estimation of 3-D ICs



A digital system essentially comprises logic, memory and interconnect. Based on hierarchical models from physical level models including the ones described here for TSV and thermal analysis as well as on-chip interconnect models and nanometer device models, to system-level architecture models that describe the organisation of the logic and memory in the 3-D stack, we have developed an analysis technique to compare the computational efficiency of 3-D integrated silicon systems for various topologies. This is available as a web-based tool here.

Link to tool: Performance Estimation

Documentation

Book chapter describing models

Help file

Conclusions

- We have created a set of models to predict the computational performance of 2-D and 3-D silicon systems.
- 3-D systems can attain up to 20-30% greater computational efficiency for every doubling of the stack.
- Moving DRAM into the stack enables over an order of magnitude savings in the interconnect energy.
- The same performance with the same power can be realized in 3-D topologies with much smaller area and at lower frequency.
- The models can be used to provide an early estimate of the performance limitations and capabilities of various processing systems before fine-grained layout and technological details are known.

Thank You!

- **European Union research funding under grant FP7-ICT-215030 (ELITE) is gratefully acknowledged.**
  - **Numonyx (Micron)** in Italy
  - **CEA-LETI** in France
  - **Hyperstone** in Germany
  - **KTH** in Sweden
  - **U Lancaster**, UK