

# Promises and Limitations of 3-D Integration

Axel Jantsch, Matthew Grange, Dinesh Pamunuwa

November 29, 2010

# ICE - Intrinsic Computational Efficiency

$ICE$  = Number of 32 bit additions per Joule  
= Number of 32 bit operations per second per Watt

$$ICE^{130} = 1/(6.9 \text{ pJ}) = 144 \text{ GOPS/W}$$

$$ICE^{50} = 1/(1.85 \text{ pJ}) = 540 \text{ GOPS/W}$$

# ICE - Intrinsic Computational Efficiency

$ICE$  = Number of 32 bit additions per Joule  
= Number of 32 bit operations per second per Watt

$$ICE^{130} = 1/(6.9 \text{ pJ}) = 144 \text{ GOPS/W}$$

$$ICE^{50} = 1/(1.85 \text{ pJ}) = 540 \text{ GOPS/W}$$

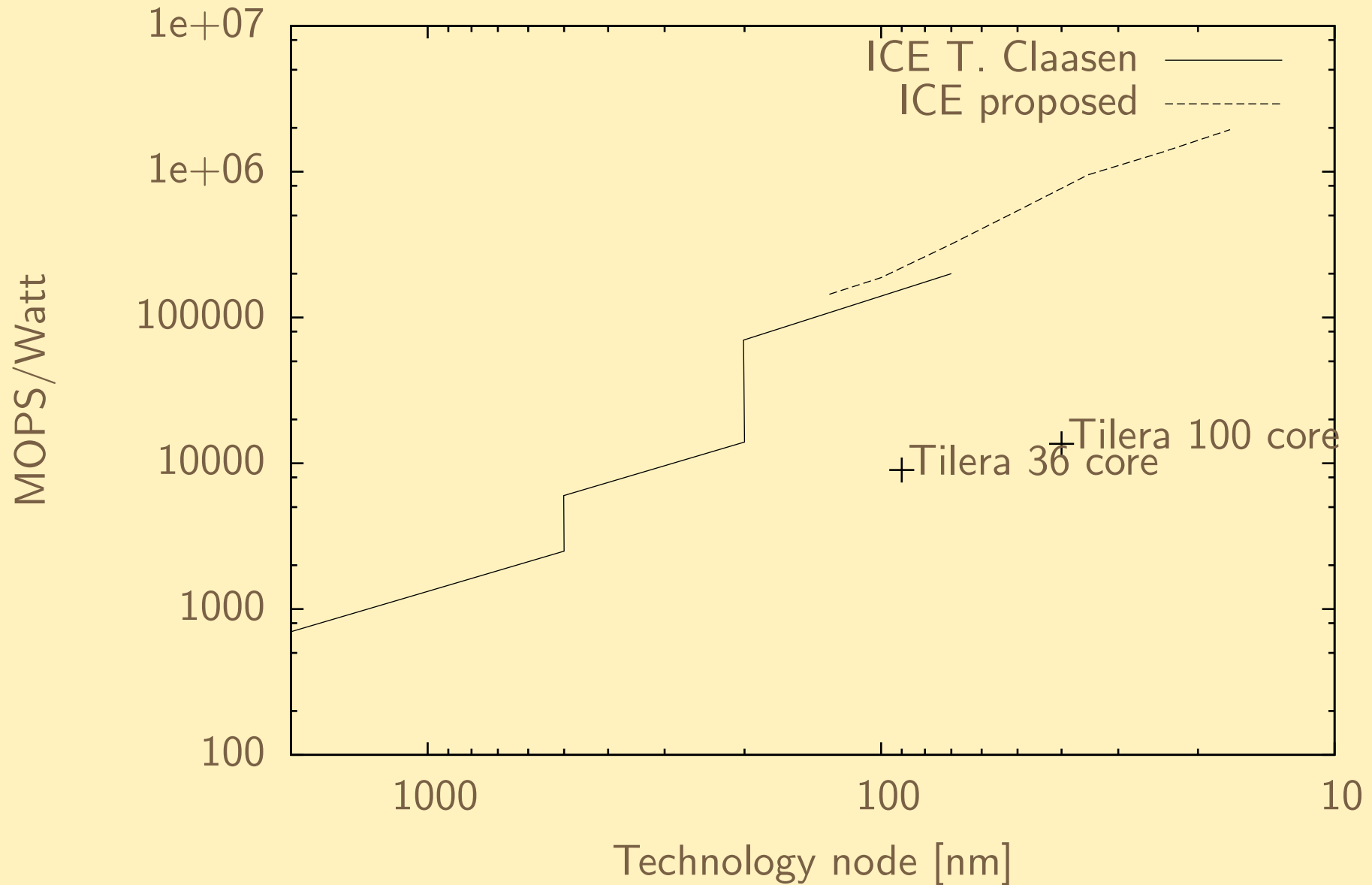
$ICD$  = Number of 32 bit adders per  $\text{mm}^2$

$$ICD^{130} = 1/(2956 \text{ } \mu\text{m}^2) = 338.3 \frac{\text{operators}}{\text{mm}^2}$$

$$ICD^{50} = 1/(437.3 \text{ } \mu\text{m}^2) = 2286.8 \frac{\text{operators}}{\text{mm}^2}$$

# The Intrinsic Computational Efficiency of Silicon

Intrinsic Computational Efficiency



# Effective Computational Efficiency

- Energy for one complete operation
- Adding memory access
- Accounting for data transportation
- Scaling of computation, memory, and wires
- Detailed models of horizontal wired and Through Silicon Vias (TSV)
- DRAM integration into 3D stack
- Considered architectures: 2D, 3D 2-, 4-, 8-, 16- layers

Not considered:

- Control
- Latency
- Local architecture variations

# ECE - Effective Computational Efficiency

$$ECE = \frac{1}{EE}$$

$$EE_{\text{arch}}^{\text{tn}} = E_{32}^{\text{tn}}$$

$$+ \mu_T (\omega (e_1 + \Delta \times E_{\text{int}}^{\text{tn}}))$$

$$+ (1 - \omega) (e_1 + E_{\text{int}}^{\text{tn}} + E_{\text{offchip}})$$

operation

on-chip memory access

off-chip memory access

$\omega$	ratio of on- to off-chip memory ( $\omega = 1$ : all on-chip, $\omega = 0$ : all off-chip)
$\Delta$	memory distribution factor ( $\Delta = 1$ : all centralized, $\Delta = 0$ : all local)
$\mu_T$	number of memory accesses per h/w operation
$e_1$	energy for a 32-bit read/write to local SRAM
$E_{\text{int}}^{\text{tn}}$	interconnect energy required to transport a 32-bit word from a non-adjacent on-chip memory to the local cache
$E_{\text{offchip}}$	energy to read/write to off-chip memory

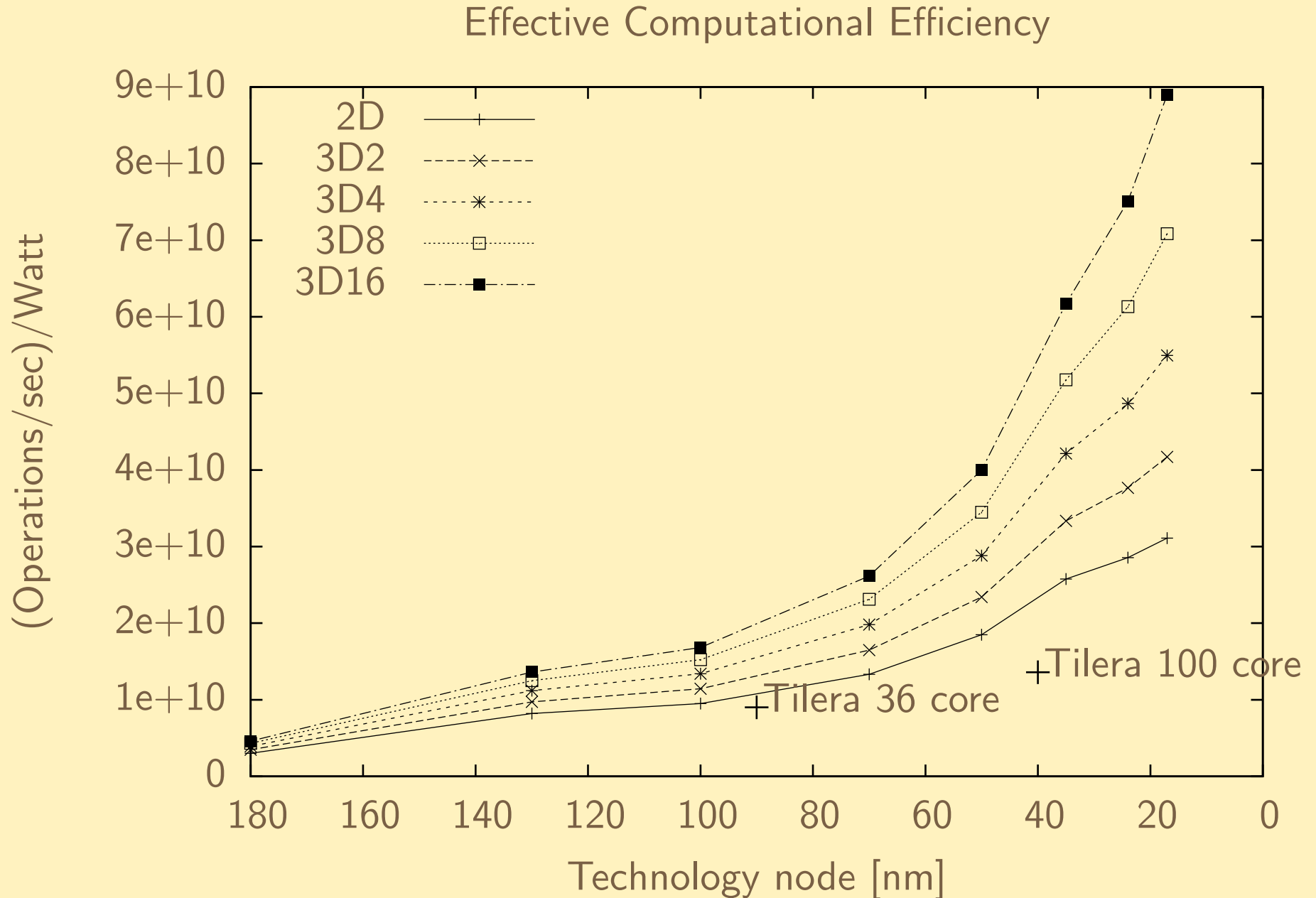
# Effective Computational Density

$$EA_{\text{arch}}^{\text{tn}} = A_{32}^{\text{tn}} + \mu_S \omega a_1 + \sigma A_{\text{int}}^{\text{tn}}_{\text{arch}}$$

operator area  
storage area  
interconnect area

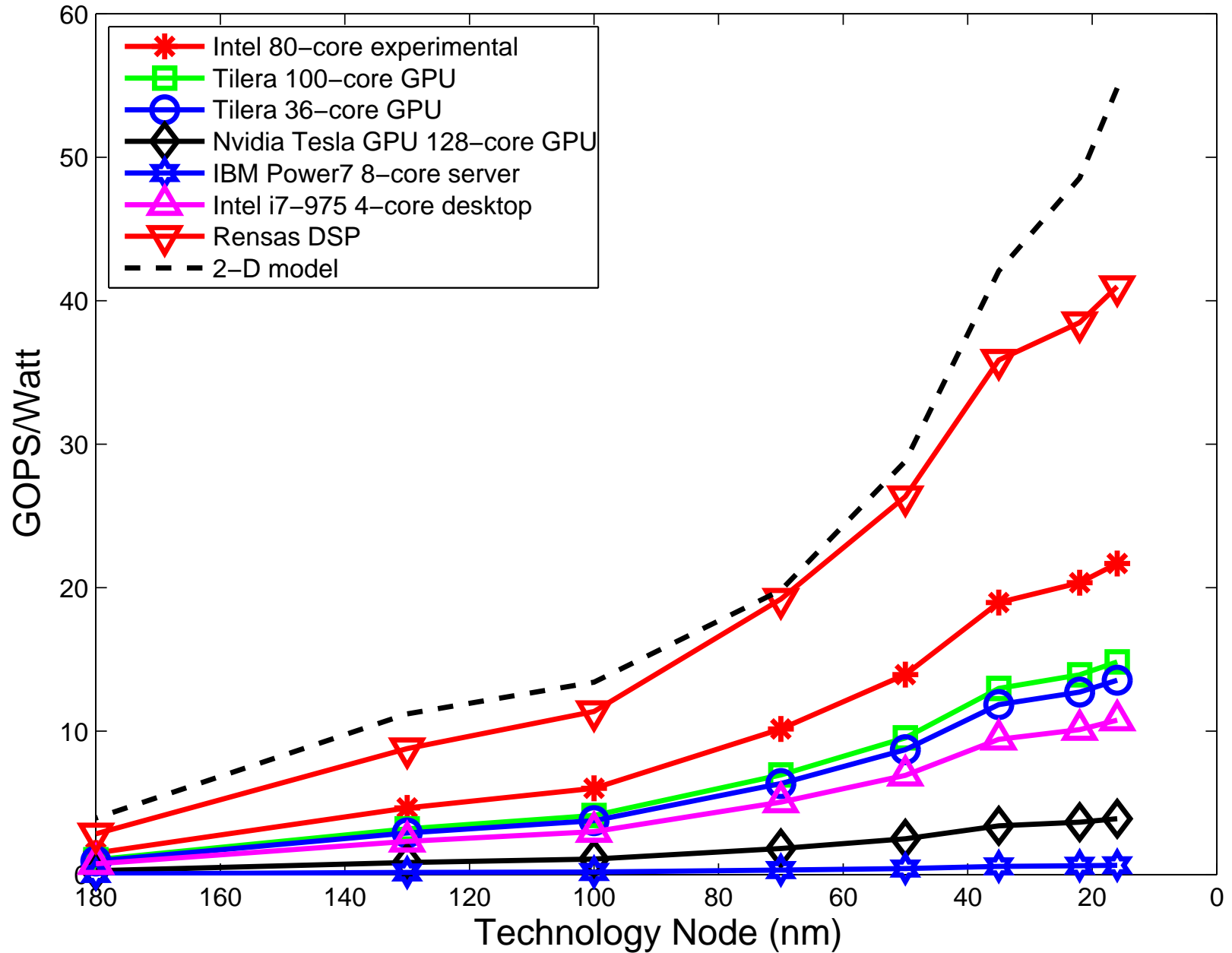
$EA_{\text{arch}}^{\text{tn}}$	Effective Area for a 32-bit addition without off-chip memory
$a_1$	area for a 32-bit memory word in SRAM or DRAM
$\mu_S$	amount of memory per operator (typically 1000-10000)
$\sigma$	interconnect sharing factor ( $\sigma = 1$ : no sharing, $\sigma = 0$ : full sharing)
$A_{\text{int}}^{\text{tn}}_{\text{arch}}$	interconnect area required to transport a 32-bit word from a non-adjacent on-chip memory to the local cache

# The Effective Computational Efficiency



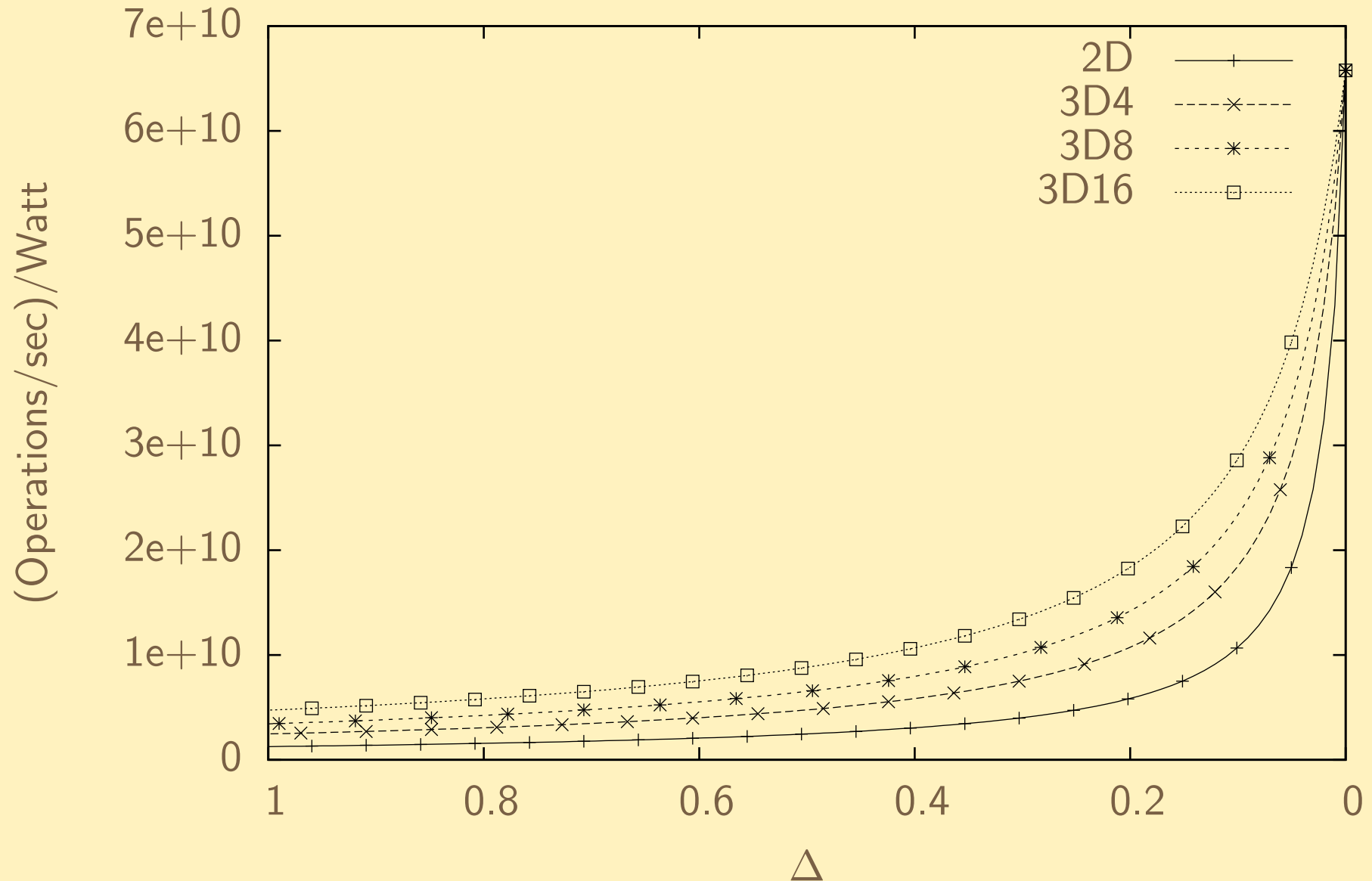


# The Effective Computational Efficiency



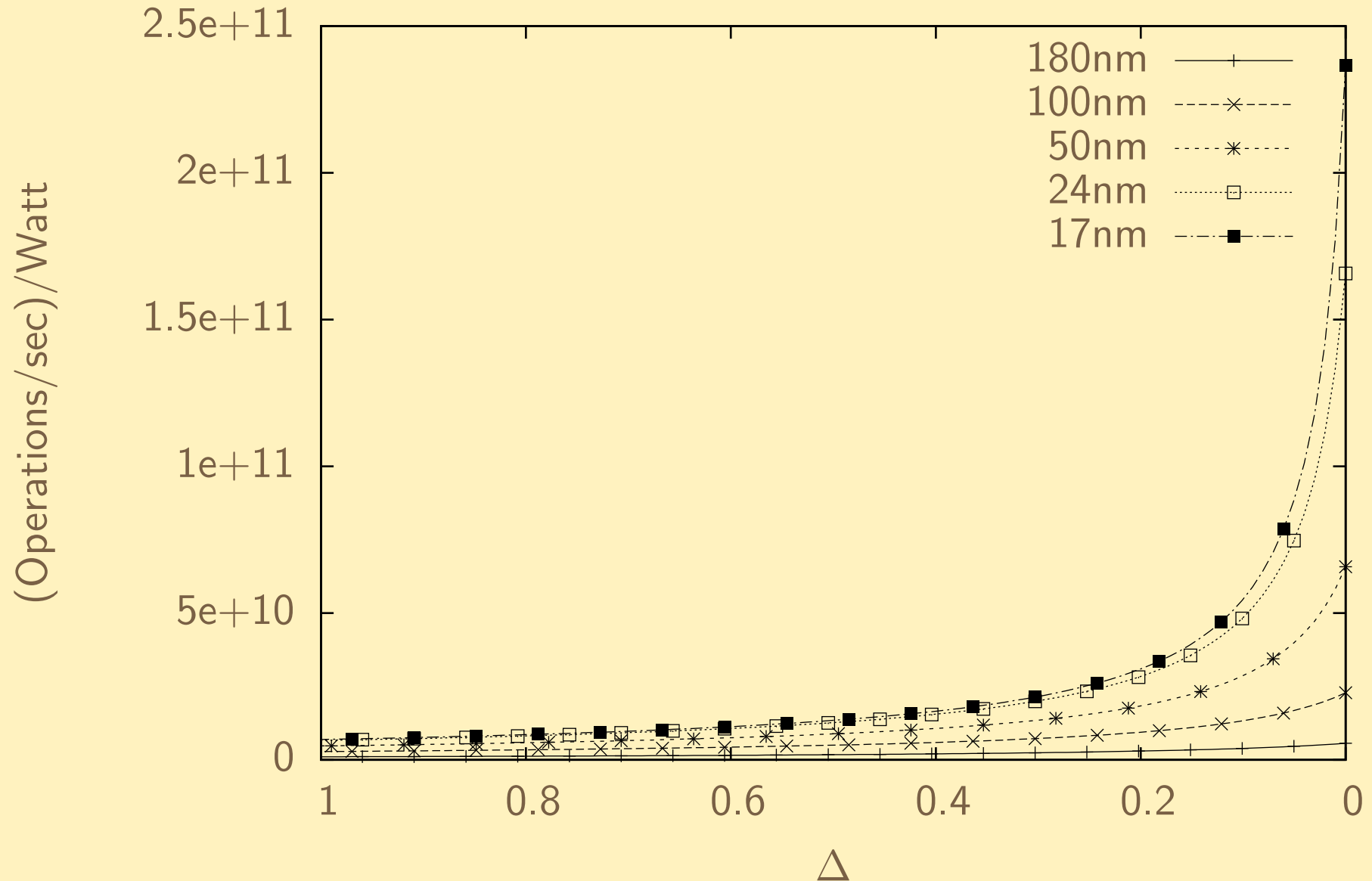
# Distributed vs Central Memory

*ECE* for  $\mu_T = 1.0, 50\text{nm}, \omega = 1.0$



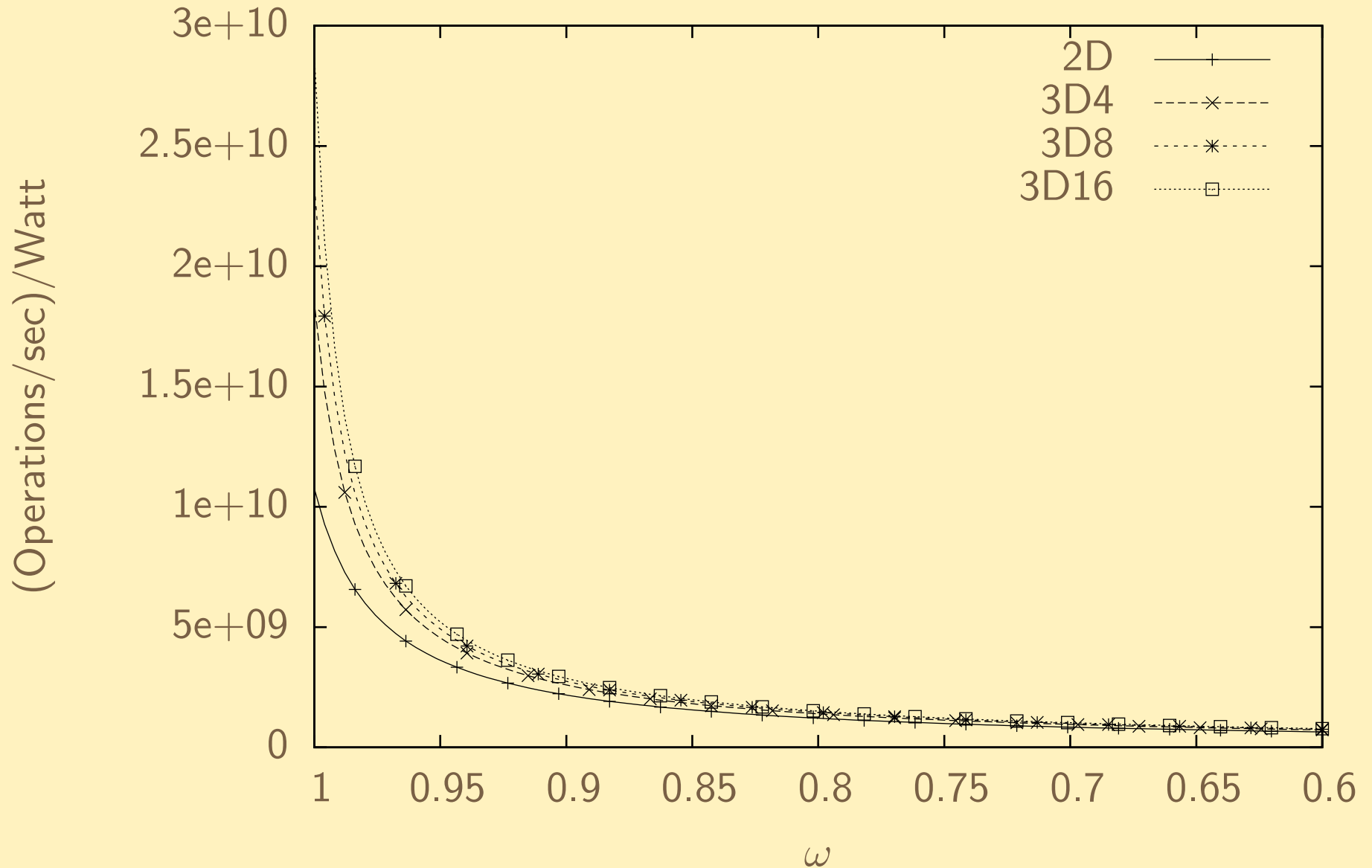
# Distributed vs Central Memory

*ECE* for  $\mu_T = 1.0, \omega = 1.0, 3D16$



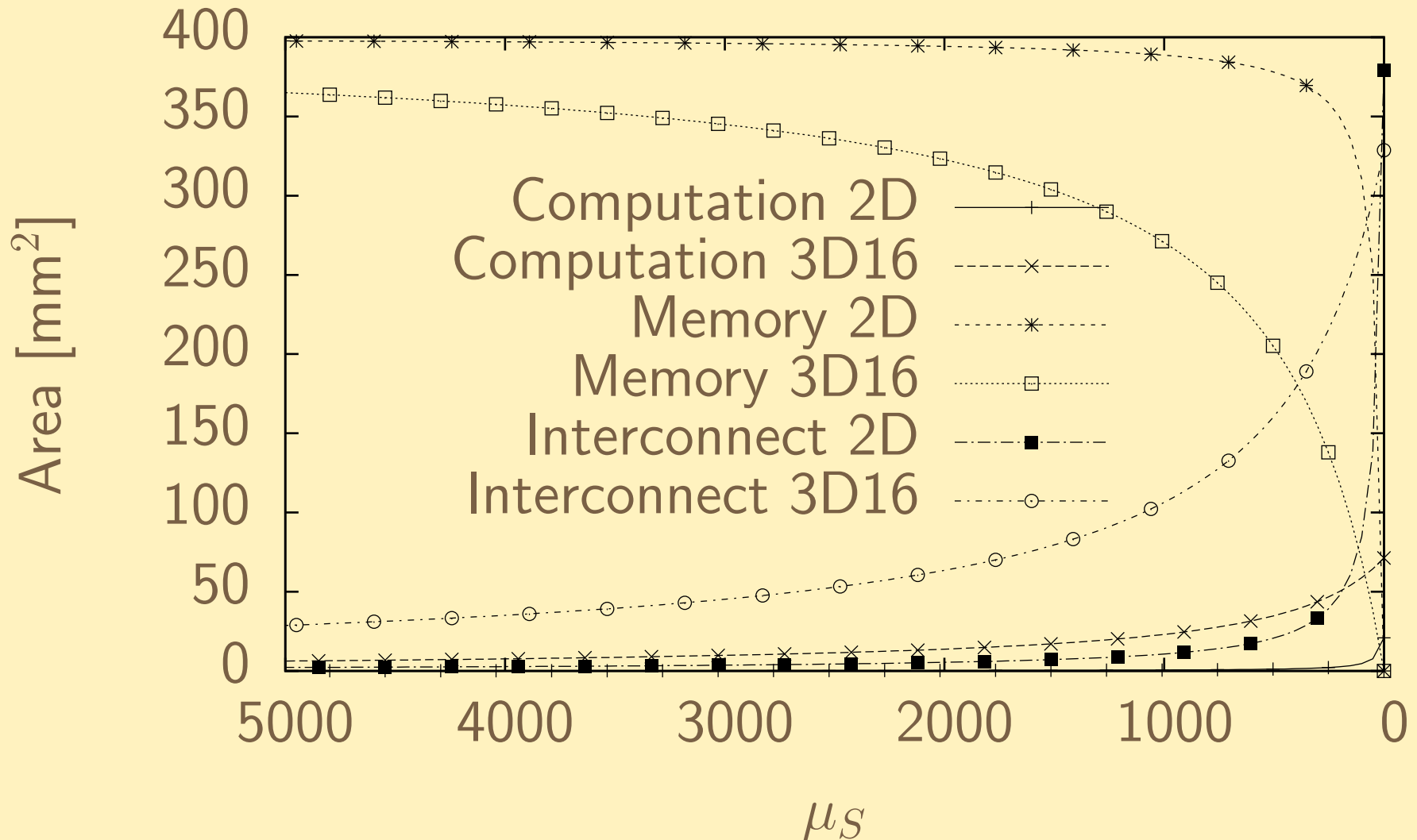
# On-Chip vs Off-Chip Memory

*ECE* for  $\mu_T = 1.0, 50\text{nm}, \Delta = 0.10$



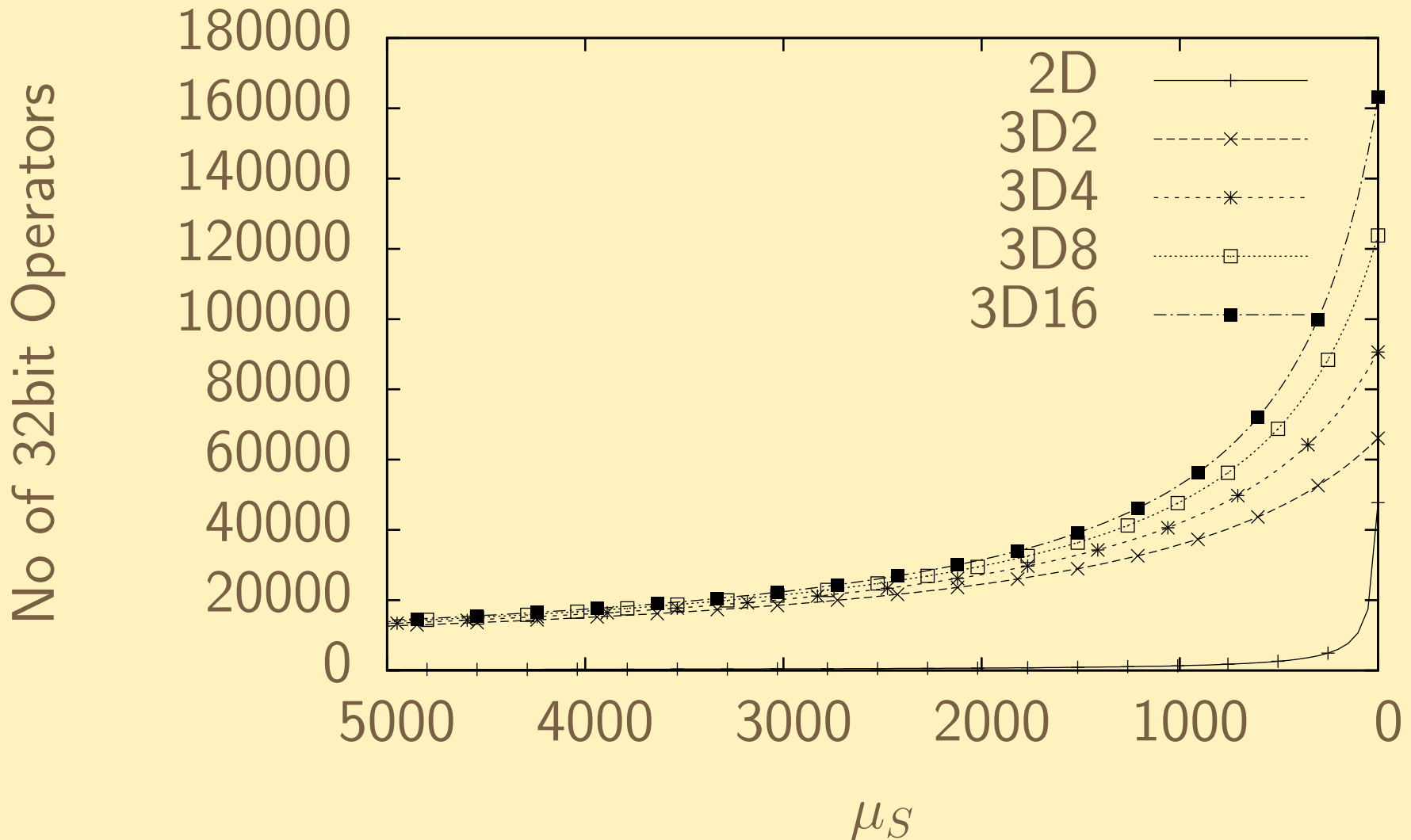
# Area Distribution for a 400mm<sup>2</sup> System

System @1Ghz, 50nm,  $\mu_T = 1.0$ ,  
 $\Delta = 0.05$ ,  $\omega = 1.0$ ,  $\sigma = 0.031$



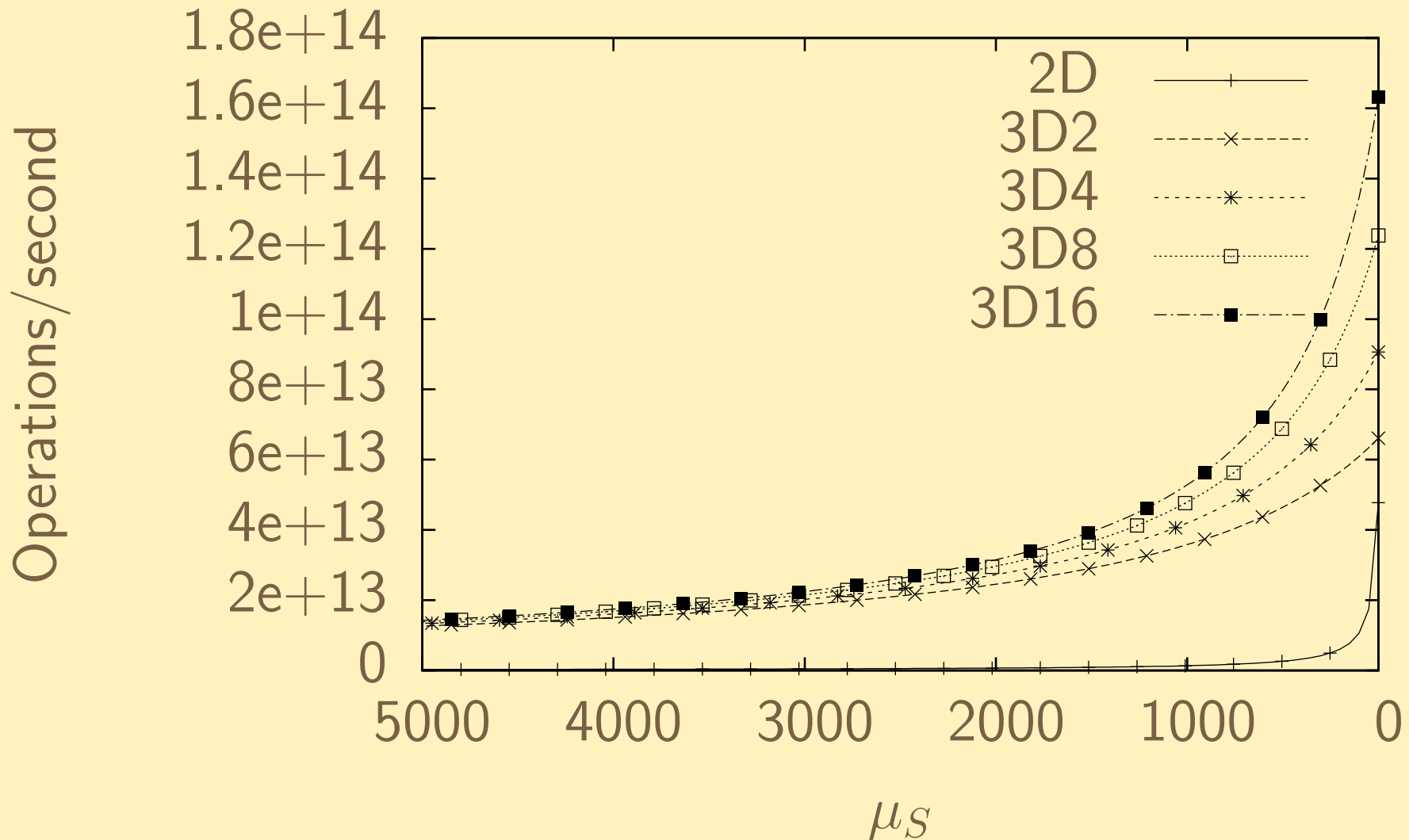
# No of Operators in a 400mm<sup>2</sup> System

System @1Ghz, 50nm,  $\mu_T = 1.0$ ,  
 $\Delta = 0.1, \omega = 1.000, \sigma = 0.031$



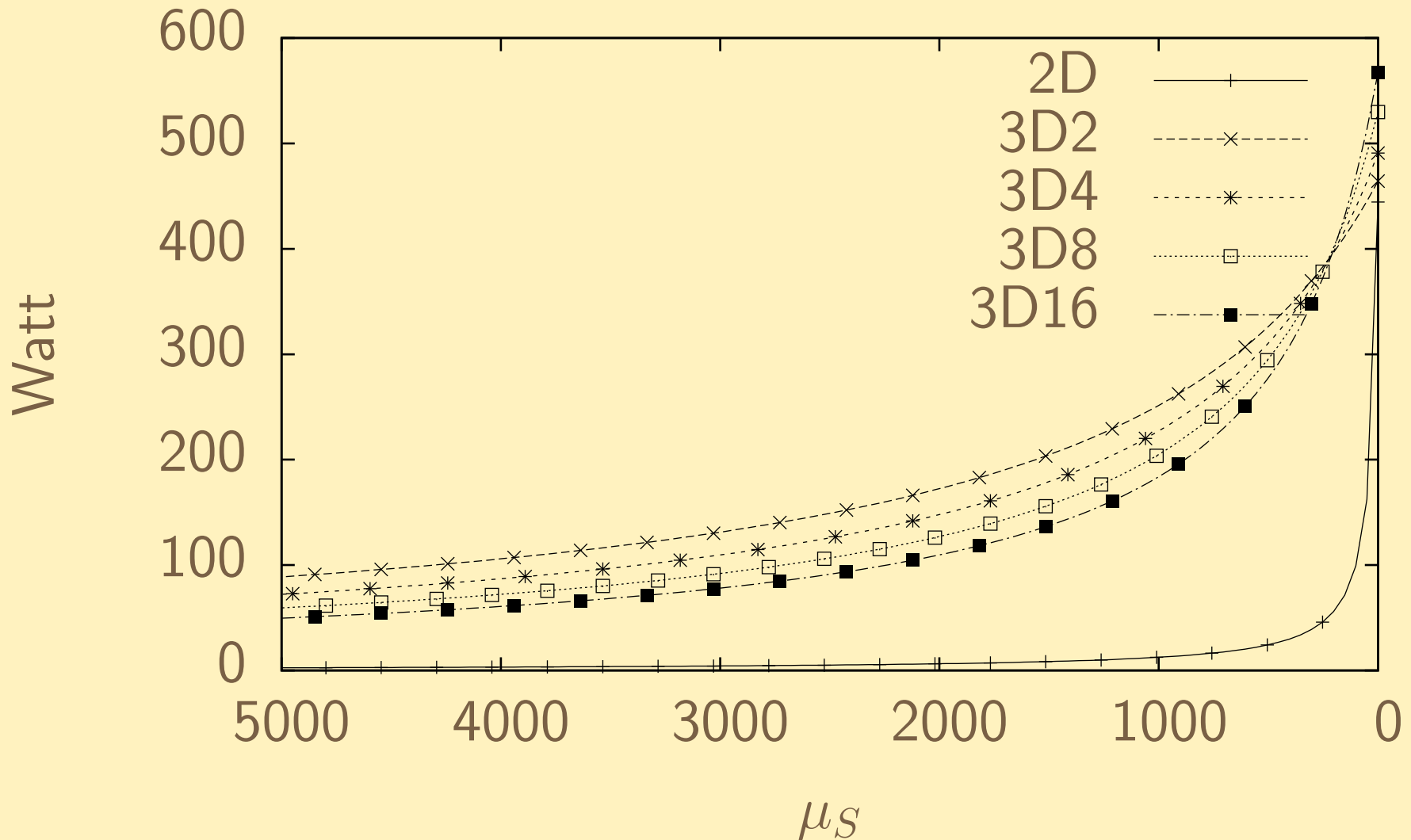
# Operations/second in a 400mm<sup>2</sup> System

System @1.0Ghz, 50nm,  $\mu_T = 1.0$ ,  
 $\Delta = 0.1, \omega = 1.0, \sigma = 0.031$



# Power Consumption in a 400mm<sup>2</sup> System

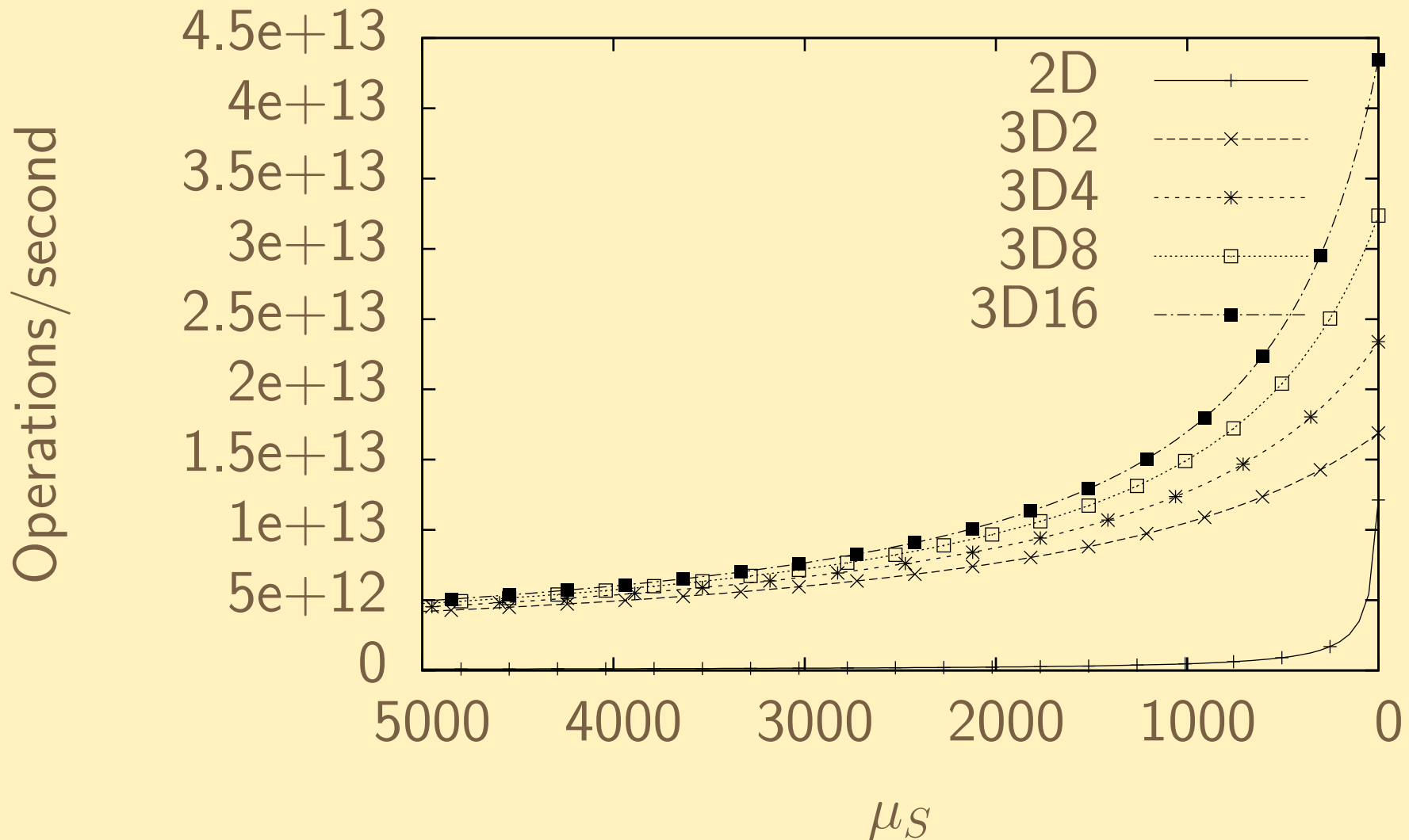
System @100Mhz, 50nm,  $\mu_T = 1.0$ ,  
 $\Delta = 0.1, \omega = 1.0, \sigma = 0.031$





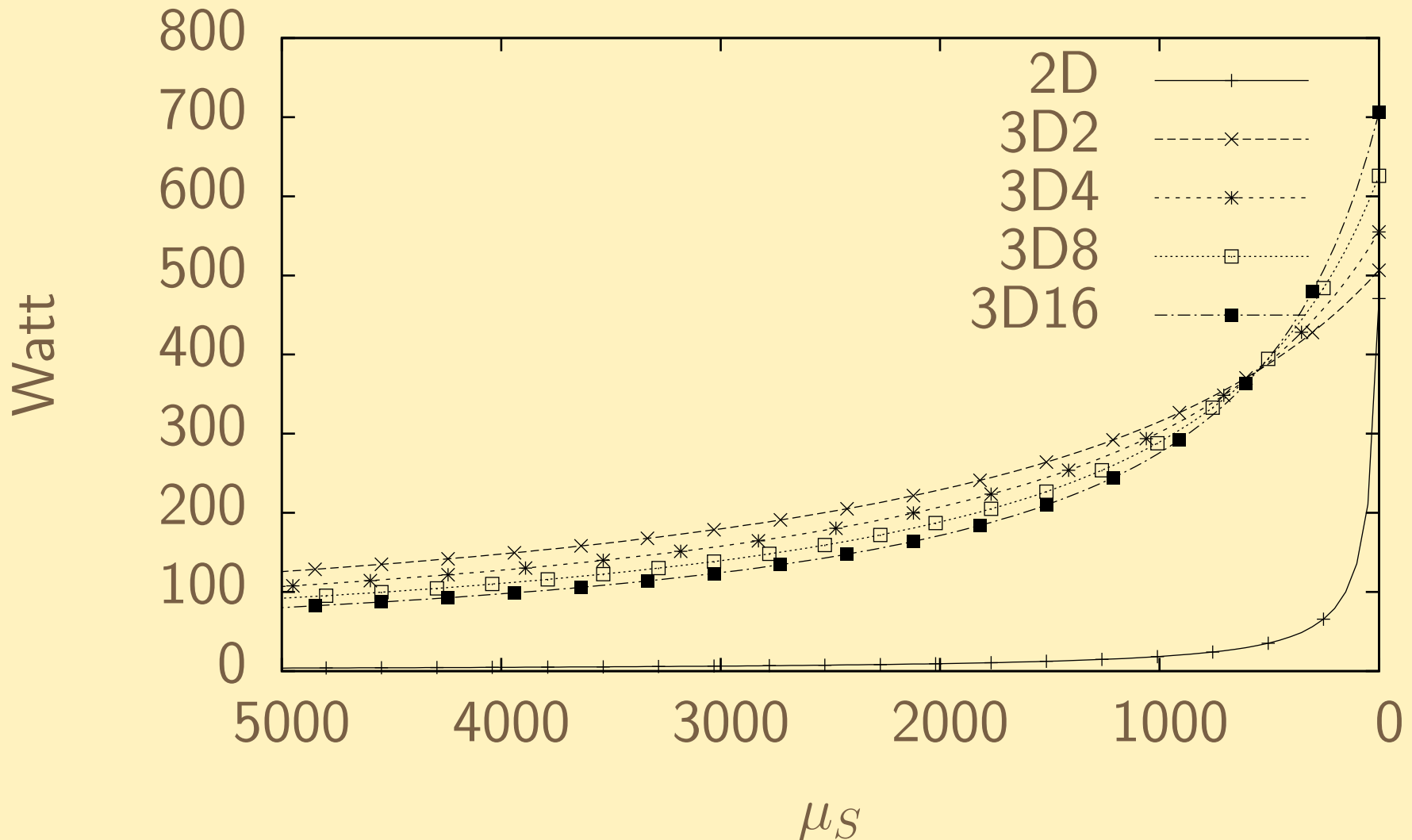
# Operations/second in a 100mm<sup>2</sup> System

System @ 700Mhz, 35nm,  $\mu_T = 1.0$ ,  
 $\Delta = 0.1, \omega = 1.0, \sigma = 0.031$



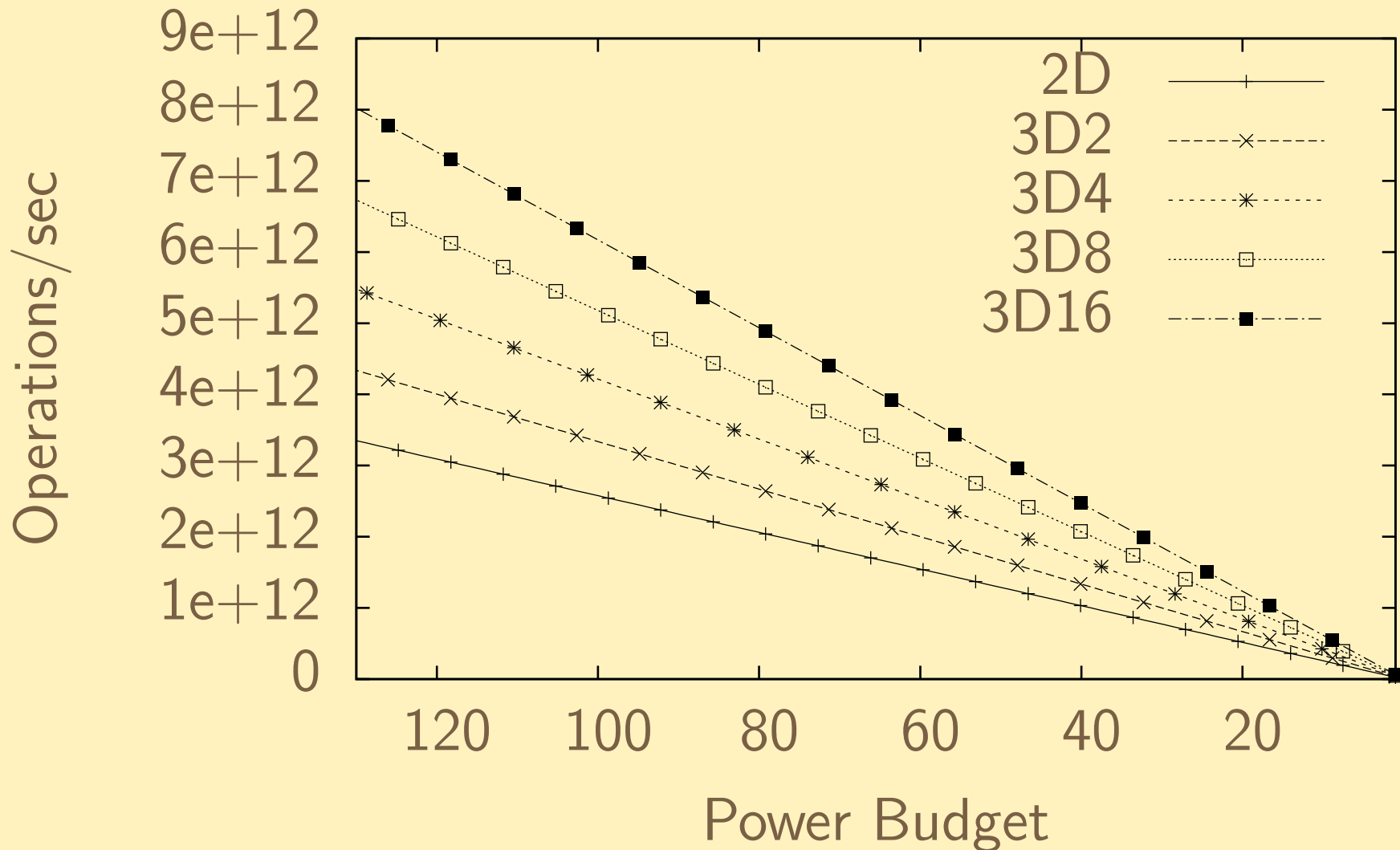
# Power Consumption in a 100mm<sup>2</sup> System

System @ 700MHz, 35nm,  $\mu_T = 1.0$ ,  
 $\Delta = 0.1, \omega = 1.0, \sigma = 0.031$



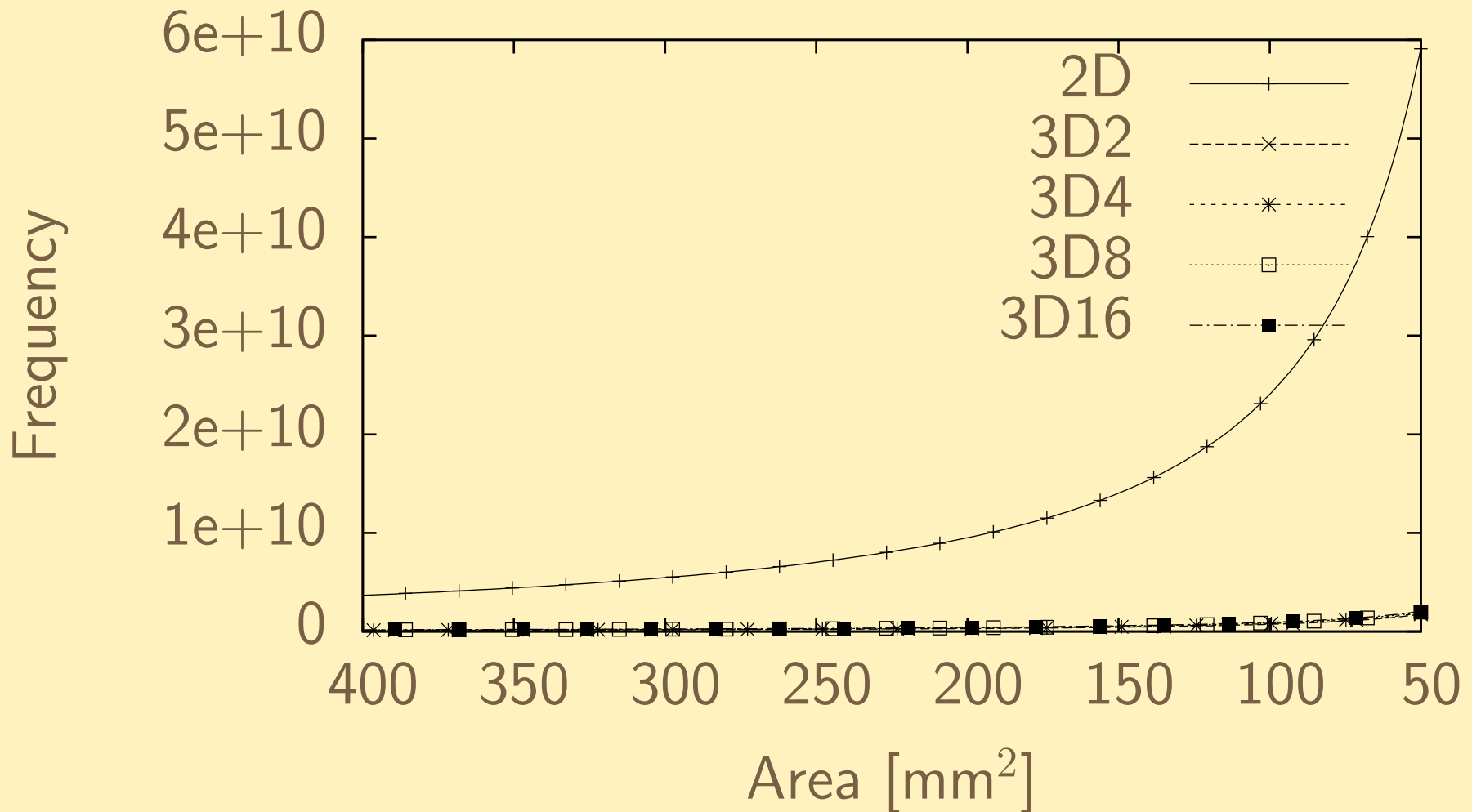
# Performance under Power Constraints

$$\mu_T = 1, \mu_S = 4000, \Delta = 0.050,$$
$$\omega = 1.0, \sigma = 0.031$$



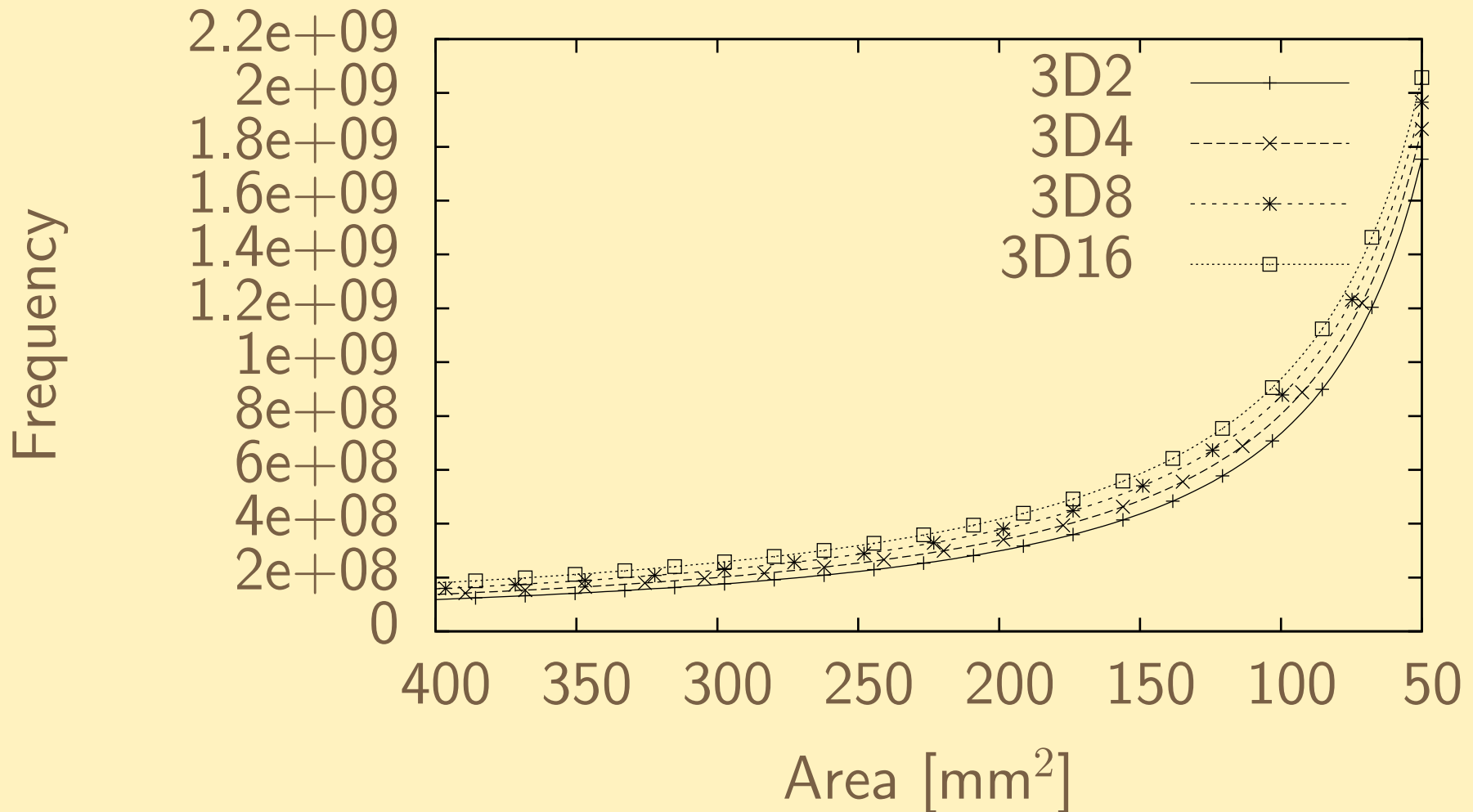
# Frequency-Area Tradeoff under Performance and Power Constraints

$\mu_T = 1.0, \mu_S = 4000, \Delta = 0.050, 35\text{nm},$   
 $\omega = 1.0, \sigma = 0.031,$   
performance=6170.5 GOPS, power=100W



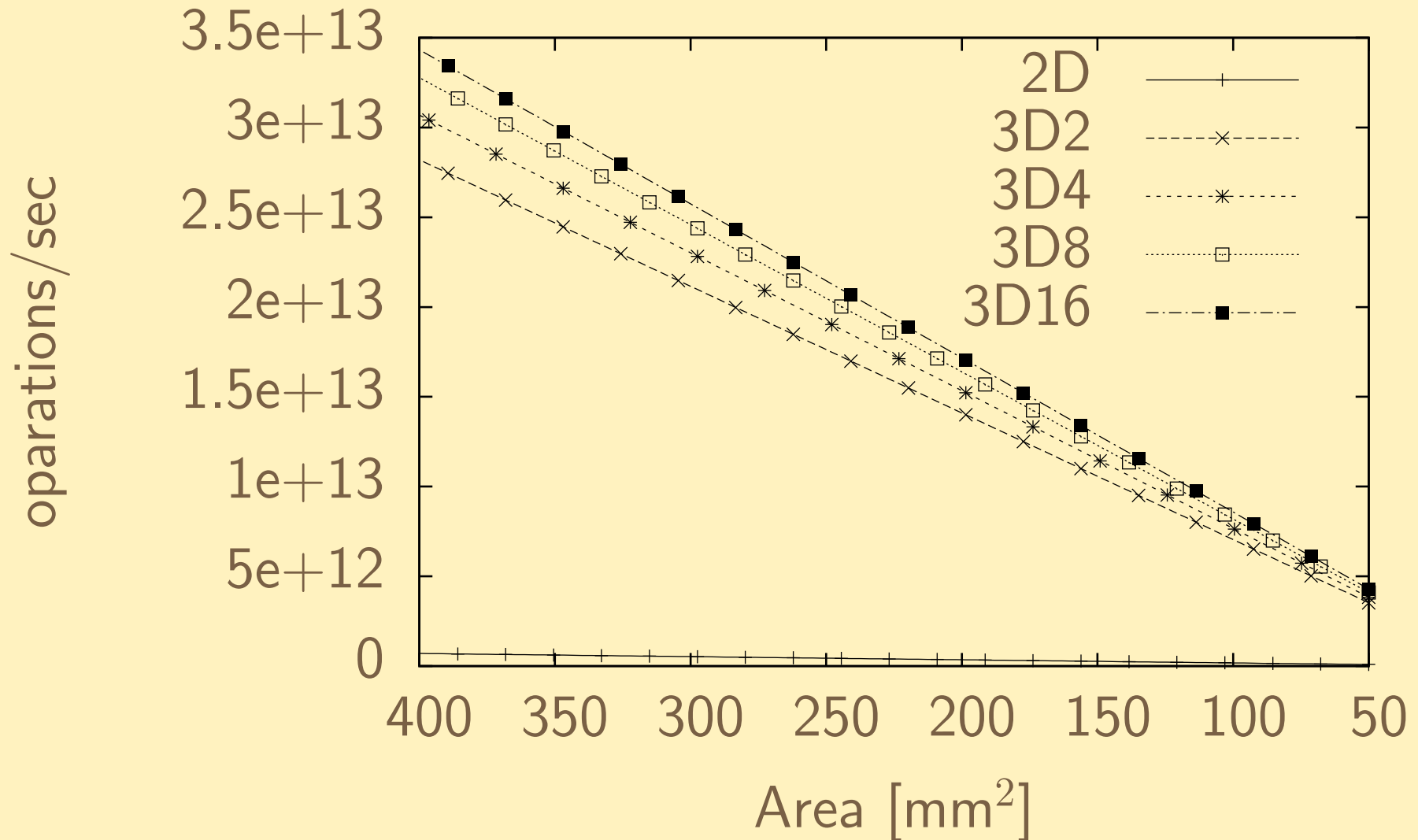
# Frequency-Area Tradeoff under Performance and Power Constraints

$\mu_T = 1.0, \mu_S = 4000, \Delta = 0.050, 35\text{nm},$   
 $\omega = 1.0, \sigma = 0.031,$   
performance=6170.5 GOPS, power=100W



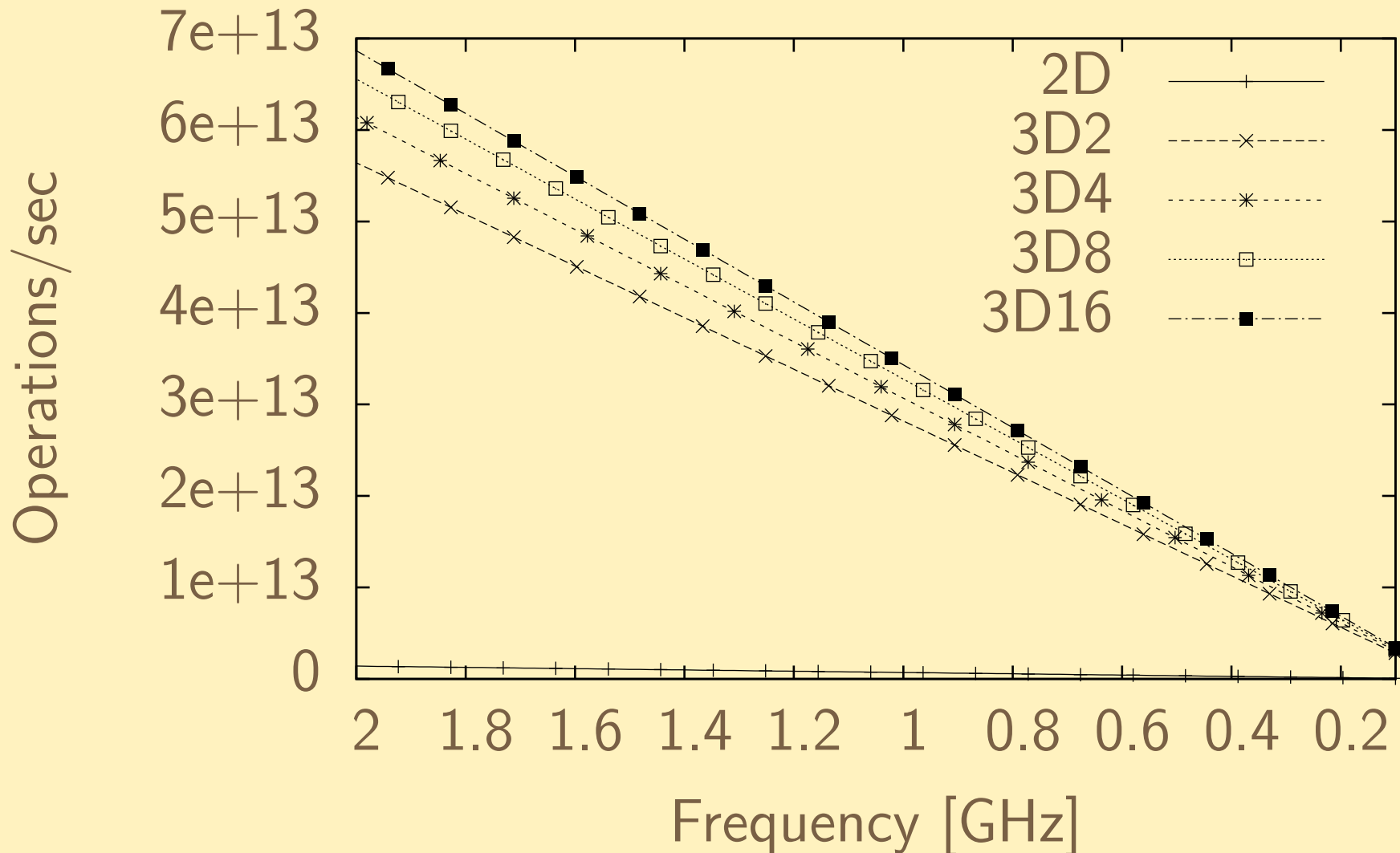
# Performance-Area Tradeoff under Power and Frequency Constraints

$\mu_T = 1.0, \mu_S = 4000, \Delta = 0.050, 35\text{nm},$   
 $\omega = 1.0, \sigma = 0.031, \text{frequency} = 1.0 \text{ GHz}$



# Performance-Frequency tradeoff under Area Constraints

$$\mu_T = 1.0, \mu_S = 4000, \Delta = 0.050, 35\text{nm},$$
$$\omega = 1.0, \sigma = 0.031, \text{area}=400 \text{ mm}^2$$



# Summary

- Study on Throughput bounds
- 3-D systems have 2 to 3 times higher *ECE* due to lower interconnect power;
- 3-D systems have one order of magnitude higher memory density due to DRAM integration;
- Consequently, 3-D systems can accommodate many more computation units in a given area and with the same amount of memory;
- This allows for much higher performance but causes also very high power density.
- The same performance with the same power can be realized in 3-D topologies with much smaller area and at lower frequency.

Not considered:

- Control
- Latency
- Local architecture variations