

# How Many Cores Can We Place on the Head of a Pin

**Peter M. Kogge**

**McCourtney Prof. of CS & Engr, Concurrent Prof. of EE  
Assoc. Dean for Research**



UNIVERSITY OF  
NOTRE DAME

# How Many Cores ~~Can~~ *Should* We Place on the Head of a Pin *- And Why?*

**Peter M. Kogge**

**McCourtney Prof. of CS & Engr, Concurrent Prof. of EE  
Assoc. Dean for Research**



UNIVERSITY OF  
NOTRE DAME

# **This Talk**

- **Given increase of “multi-core” die**
- **What are the major variations possible**
- **What are technology constraints**
- **What does this mean to chip architecture**
- **How do we optimize what’s on-chip**

**Acknowledgement: This talk is an outgrowth of one given at IWIA’05.**

# Relevant Comments from Dave Patterson

- “If memory is the problem, then ...”
- Treat cores like transistors
- Does using a large # of physical processors to support a smaller number of virtual processors make sense?
- “Flight Data Recorder”
- Transactional memory

*It's the Memory, Stupid!*

# Topics

- **How We Spend Today's Silicon**
- **Raw Performance & Storage**
- **Factoring in Overheads**
- **Chip Level Architectural Design Space**
- **Explorations**

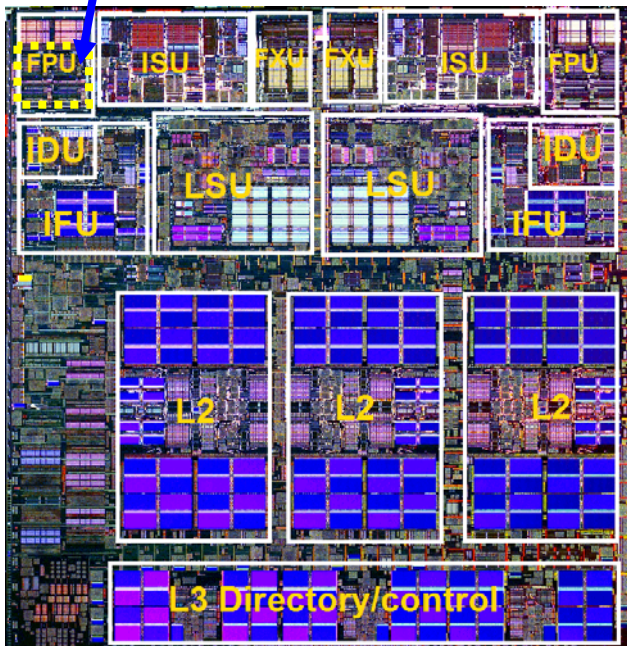
**Design Reference: a “Peta” System**

# How We Spend Our Silicon Today

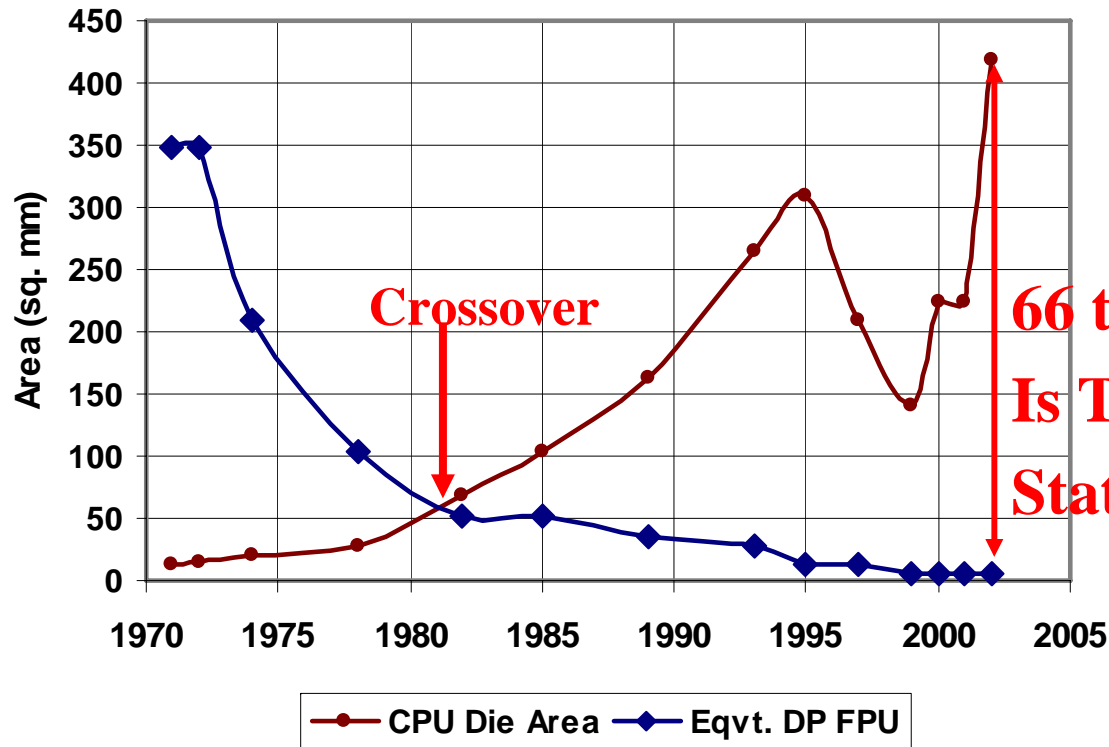
# How Are We Using Our Silicon?

## Compare CPU to a DP FPU

**IBM P5 Dual Core**

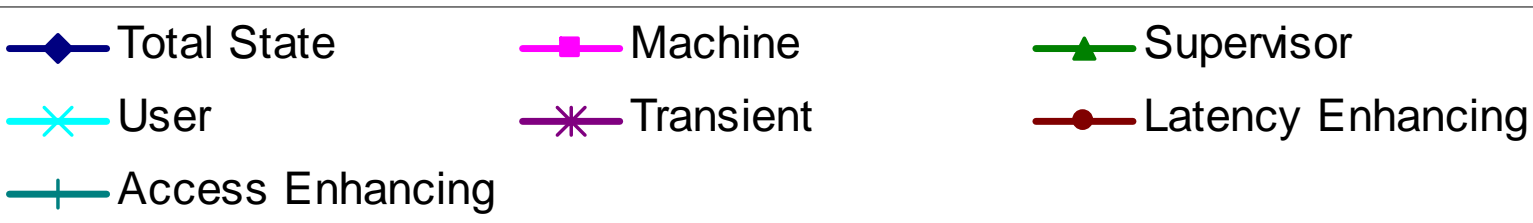
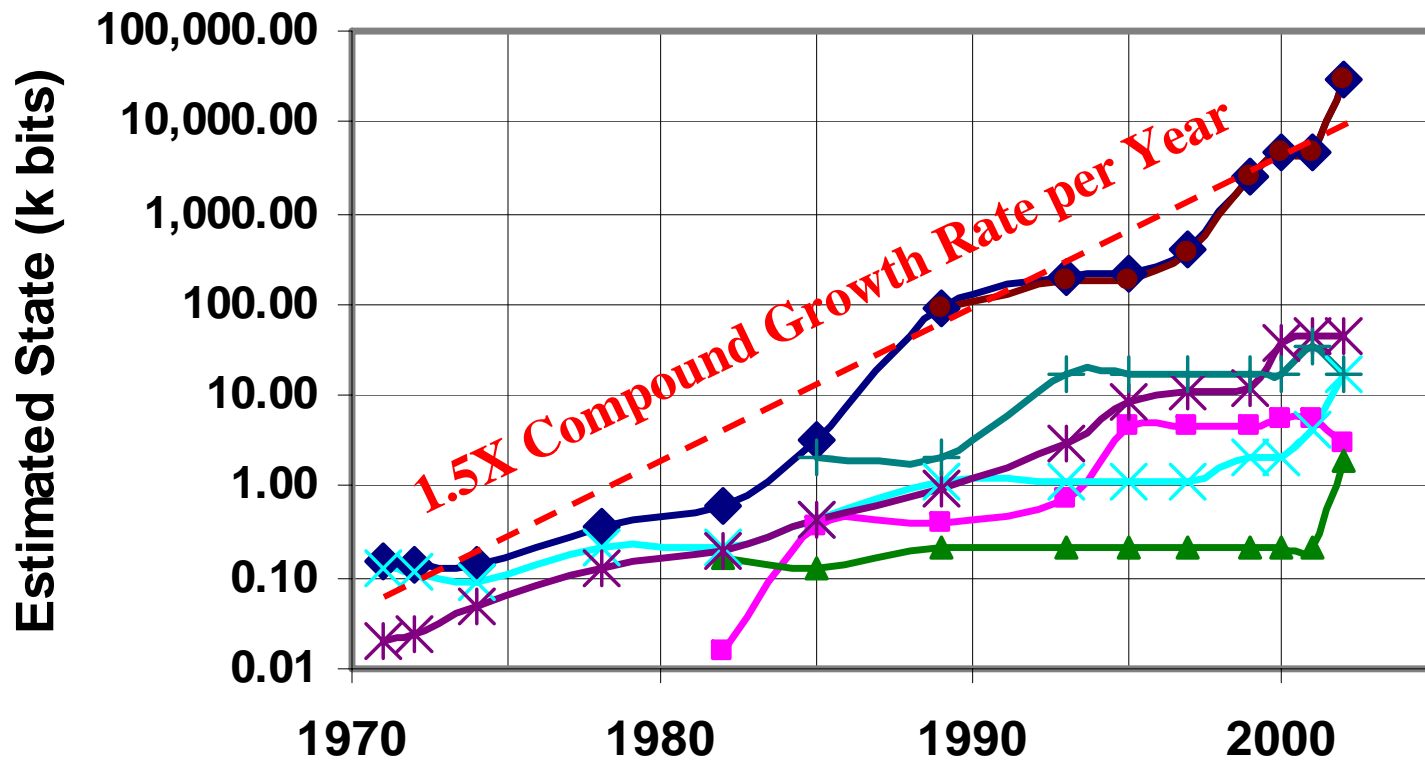


**Intel Single Core Family**

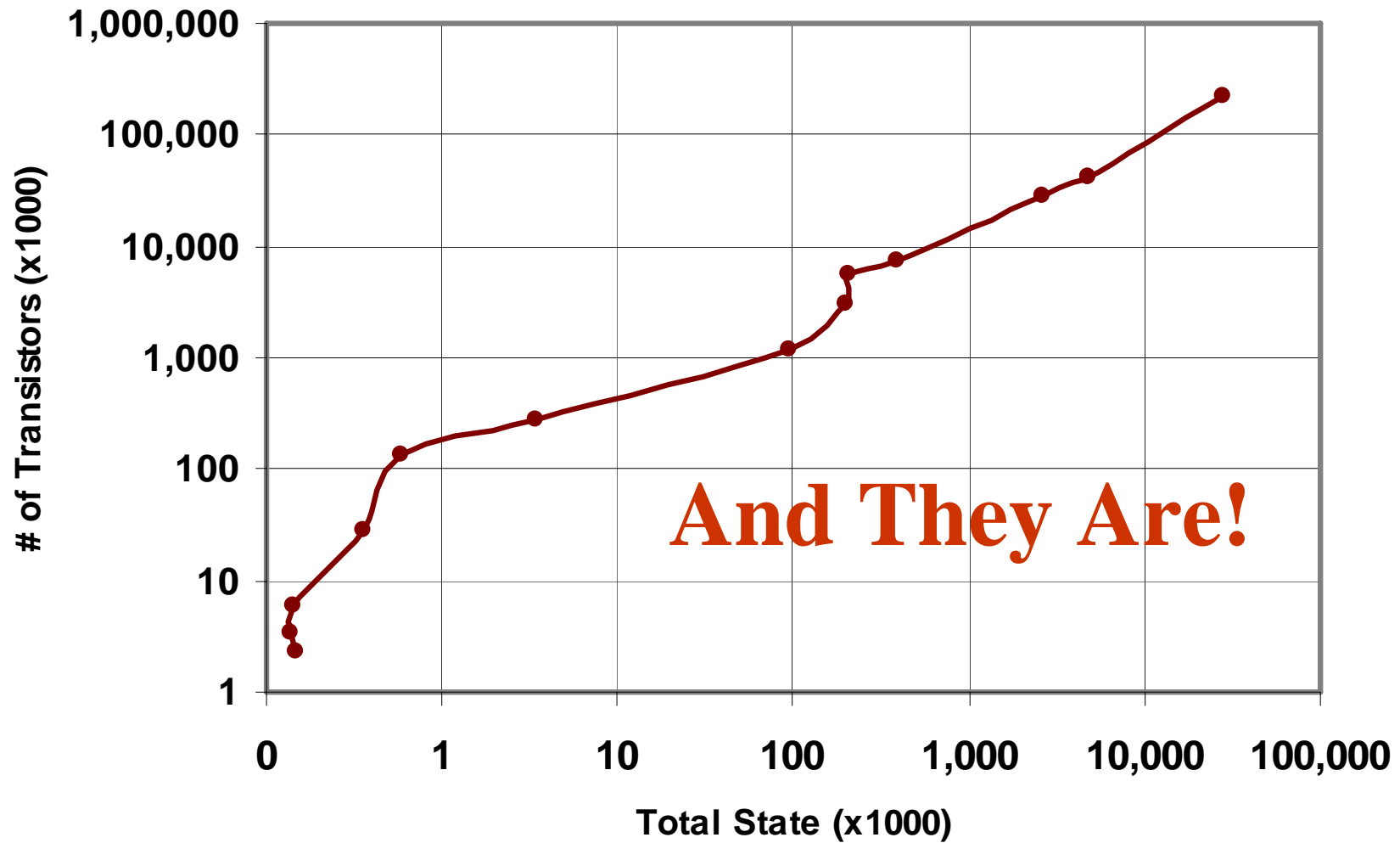


**36MB SRAM L3 chip**

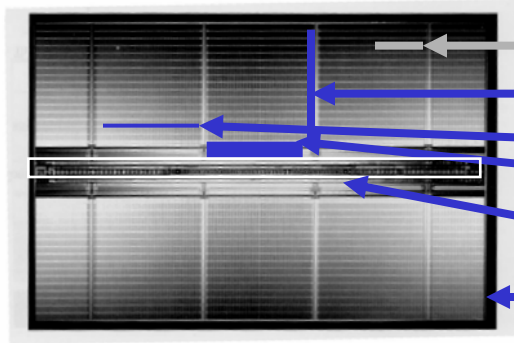
# Core CPU State vs Time



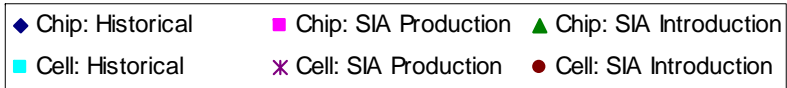
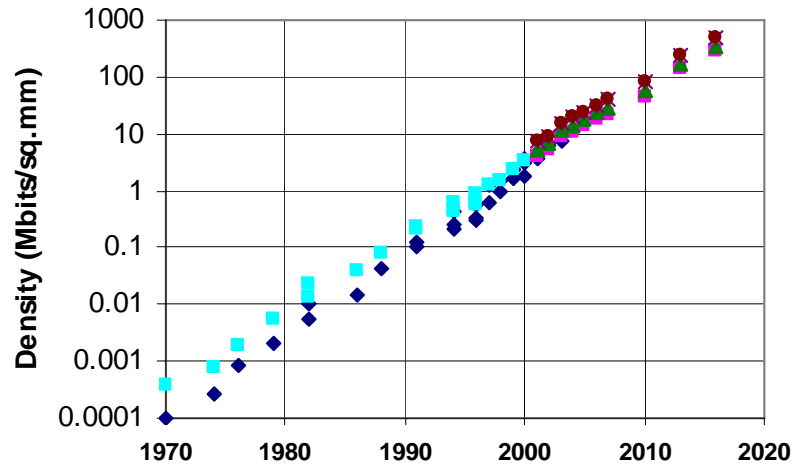
# So We Expect State & Transistor Count to be Related



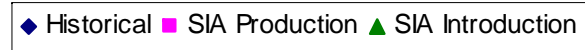
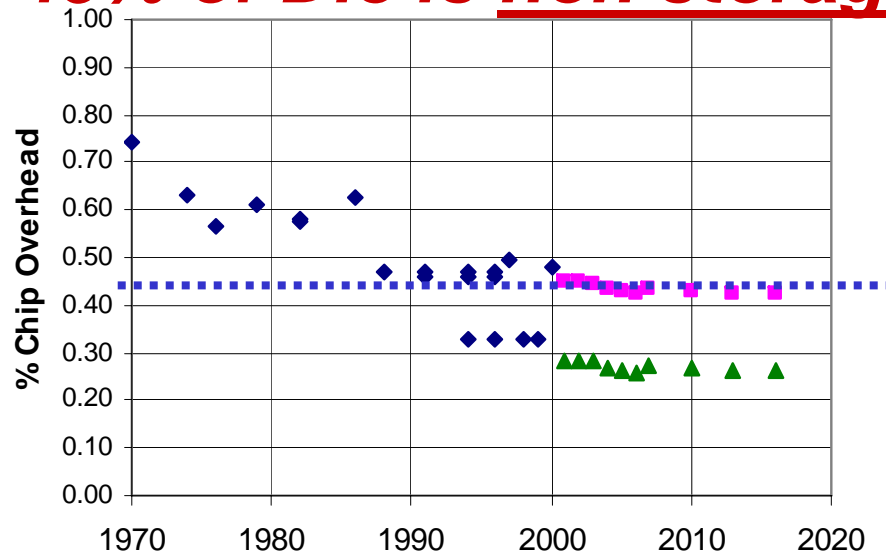
# Classical DRAM



- Memory mats: ~ 1 Mbit each
- Row Decoders
- Primary Sense Amps
- Secondary sense amps & “page” multiplexing
- Timing, BIST, Interface
- Kerf



## 45% of Die is non storage



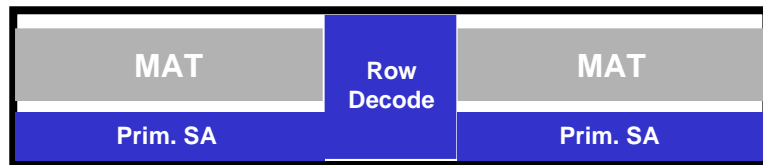
# Interesting Observation

**Our Processing Die are not all logic**

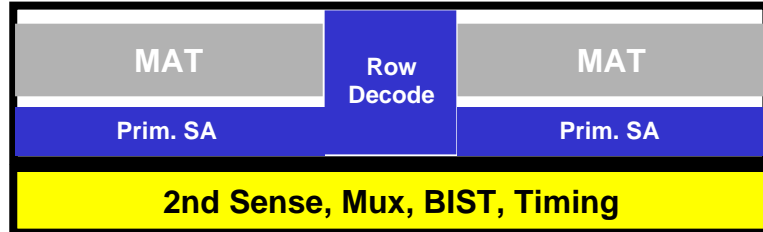
**And**

**Our Memory Die are not all storage**

# Embedded RAM Macros Today



Some maximum # of memory blocks



Memory Block:  $512 \times 2048 = 1$  Mbit

Base Block

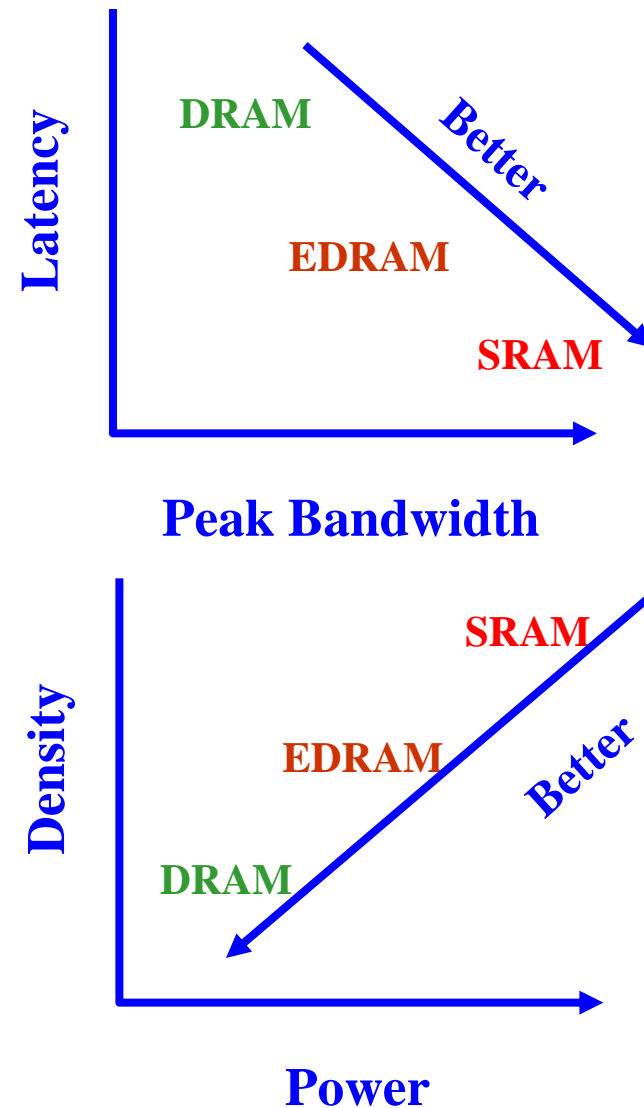
Address

“Wide Word” Data: 100’s of bits

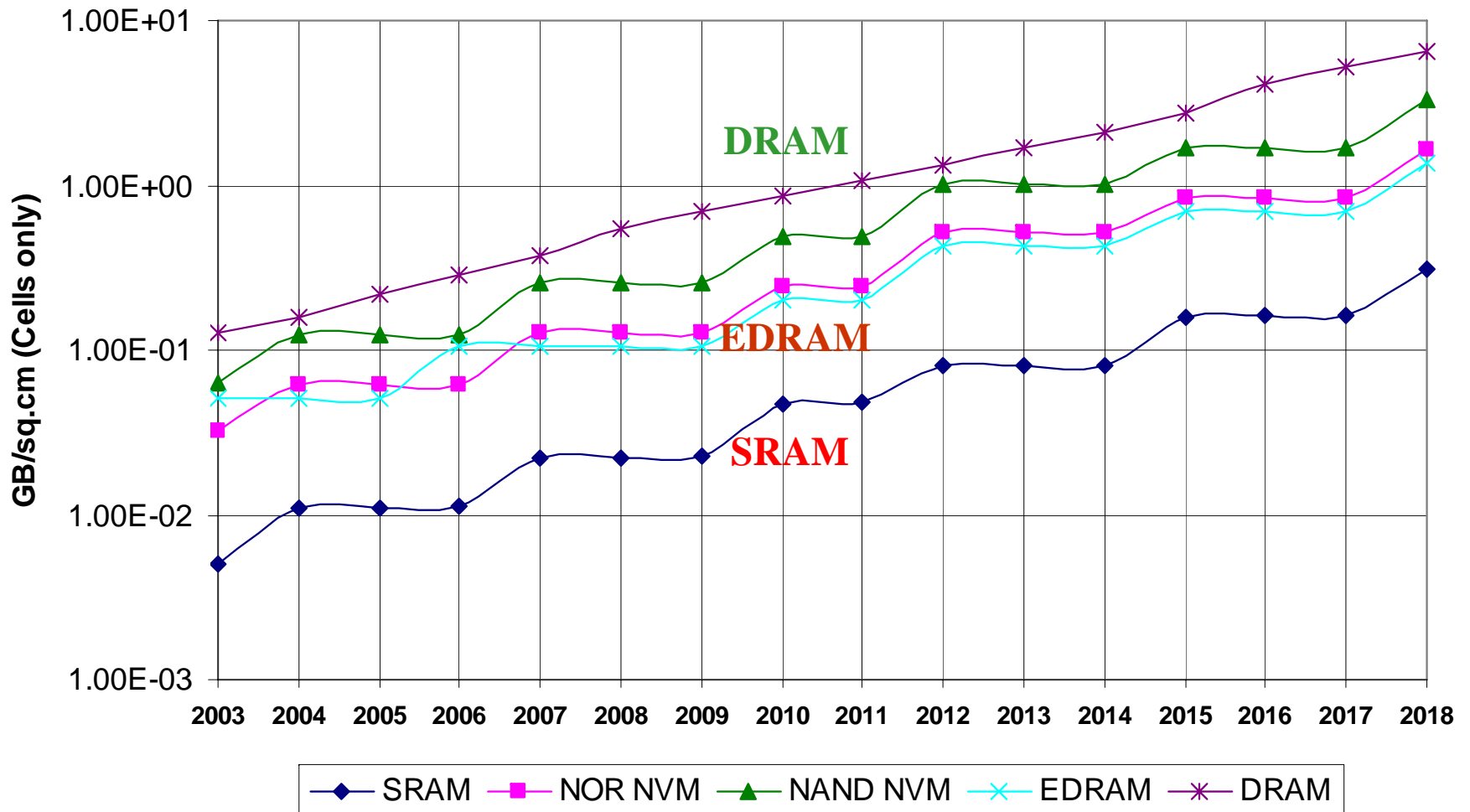
- Latencies 10 ns and below
- Peak Bandwidth/Macro into the GB/sec

# Key Types of Memory Cells

- **Commodity DRAM**
- **Embedded DRAM**
- **SRAM**
- **Non-Volatile RAM**
  - NAND Type
  - NOR Type



# Memory Storage Density: Cells Only



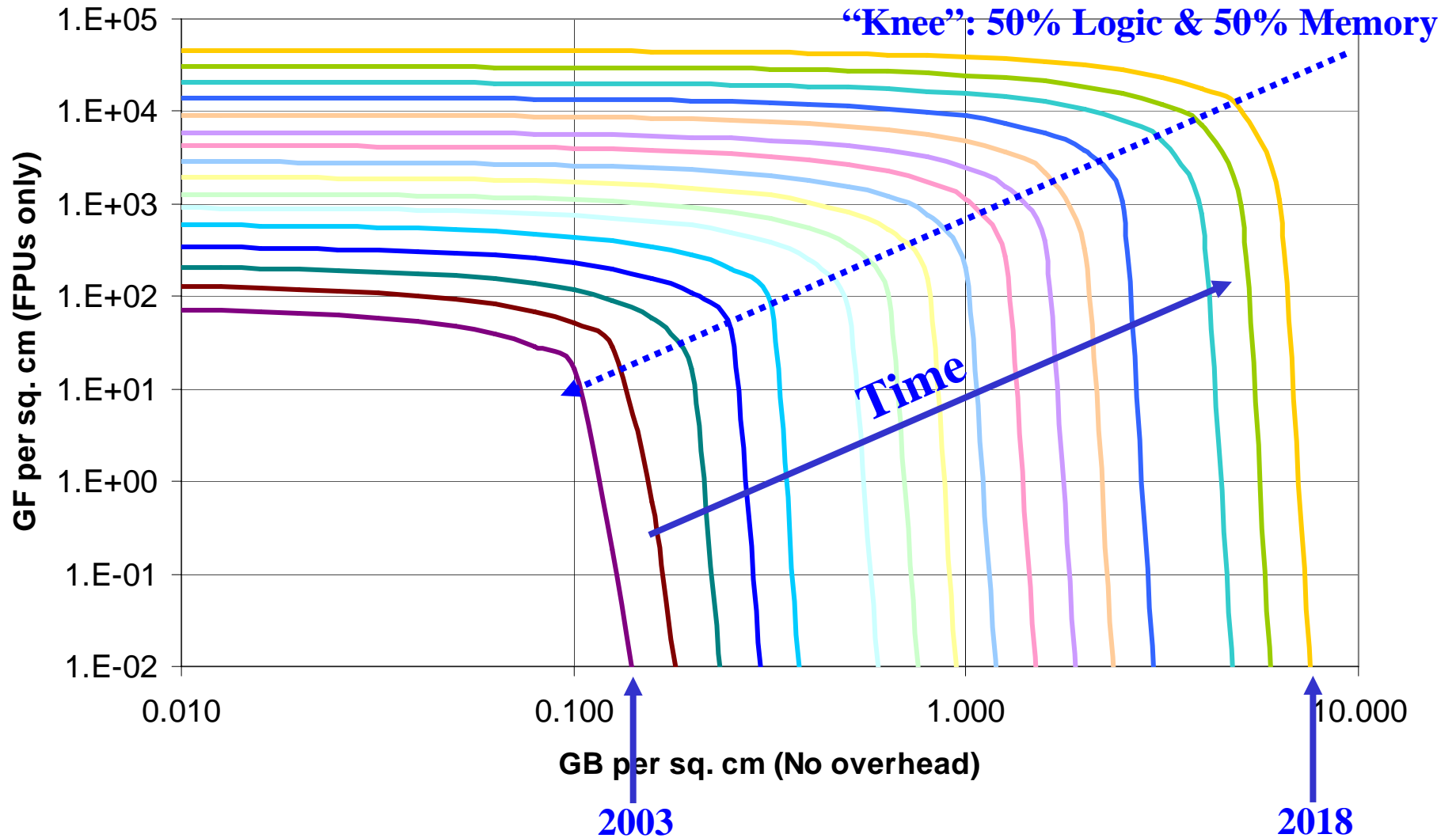
# Relative Memory Density Factors

(derived from 2004 Roadmap)

- **Commodity DRAM: 1.0**
- **NAND NVM: 1.6**
- **NOR NVM: 4.2**
- **Embedded DRAM: 4.7**
- **SRAM: 20.0**

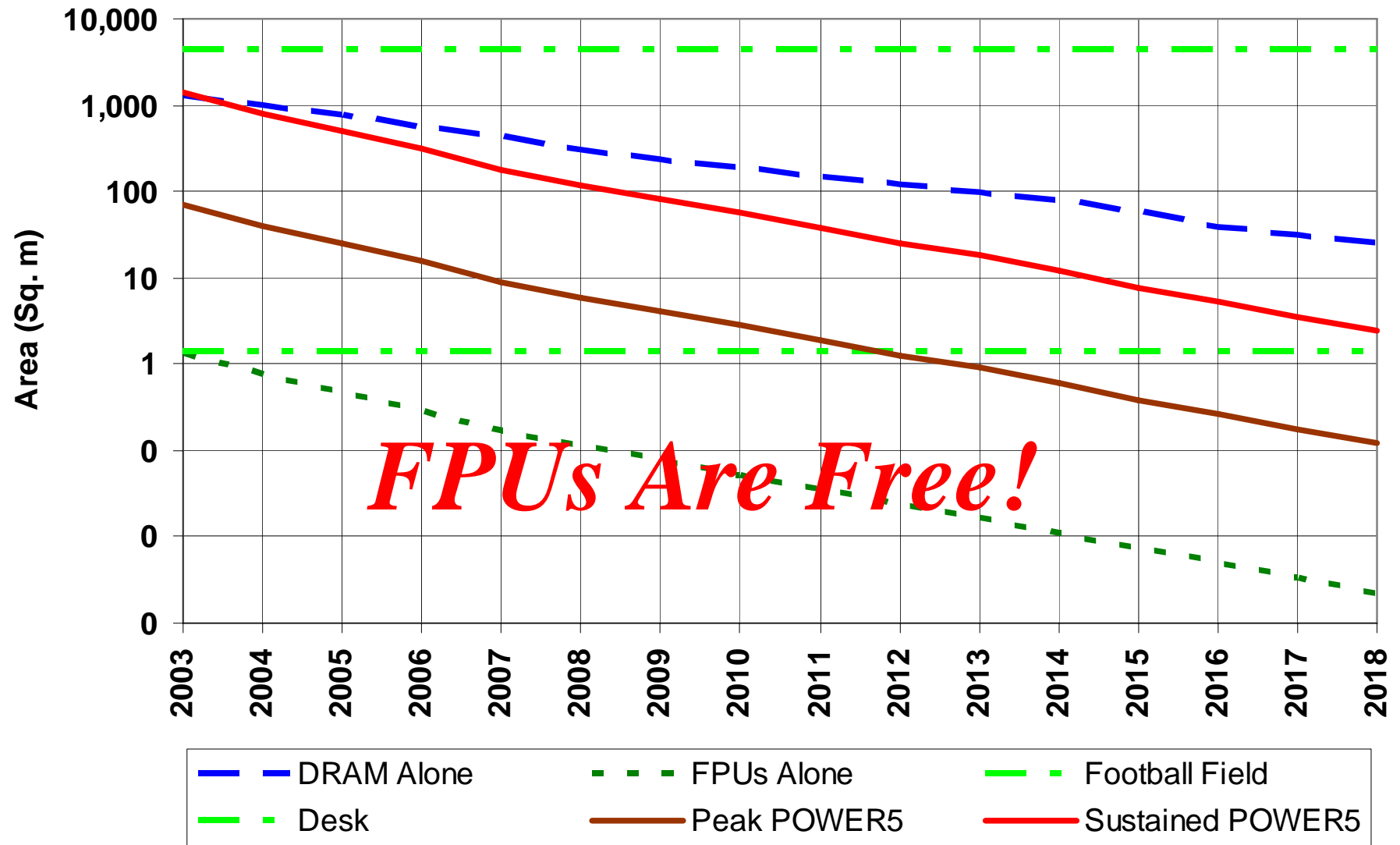
# Raw Performance & Storage

# How can we use a sq. cm? (with no overhead)





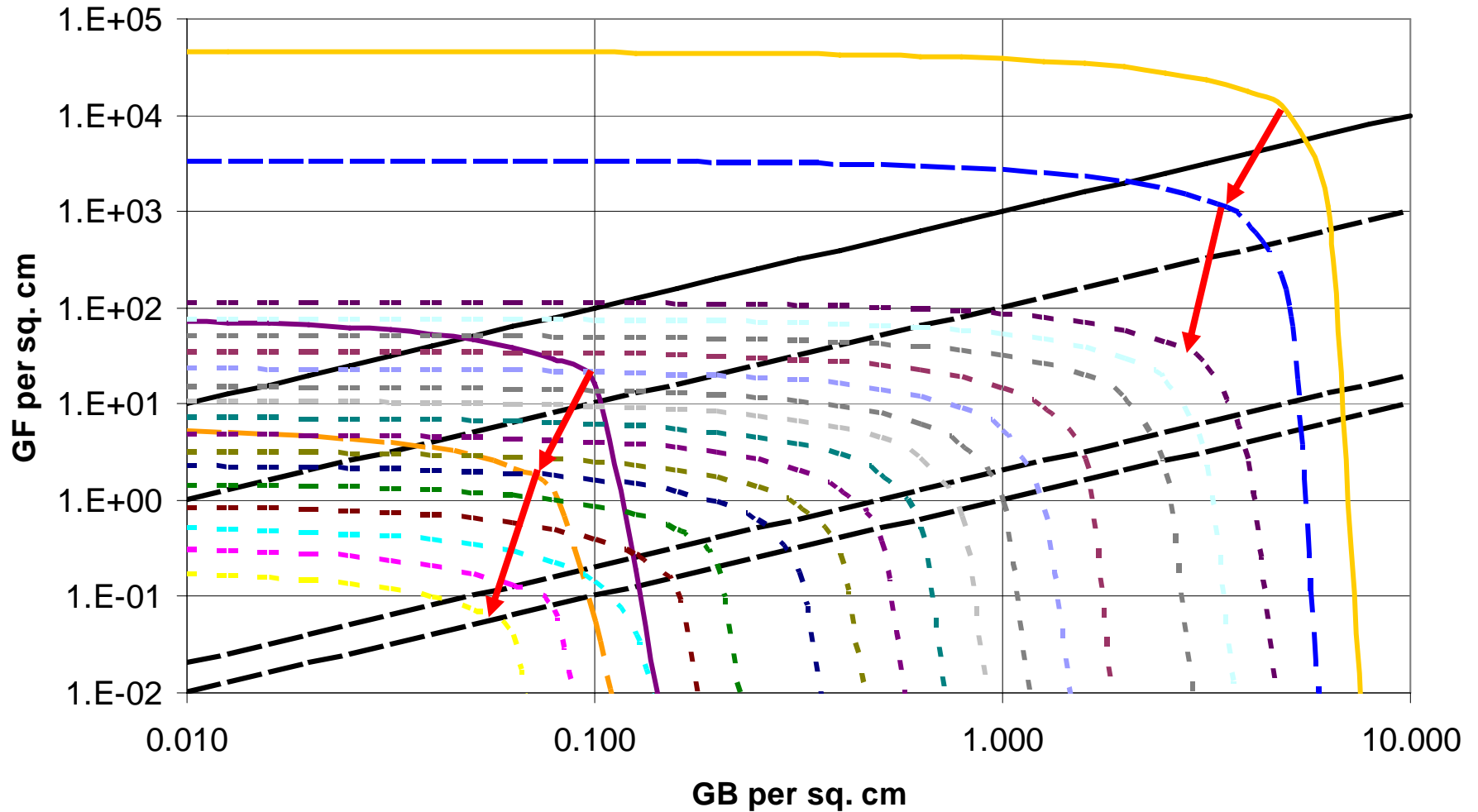
# Minimal Size for a “Peta” System



# Factoring in Overheads



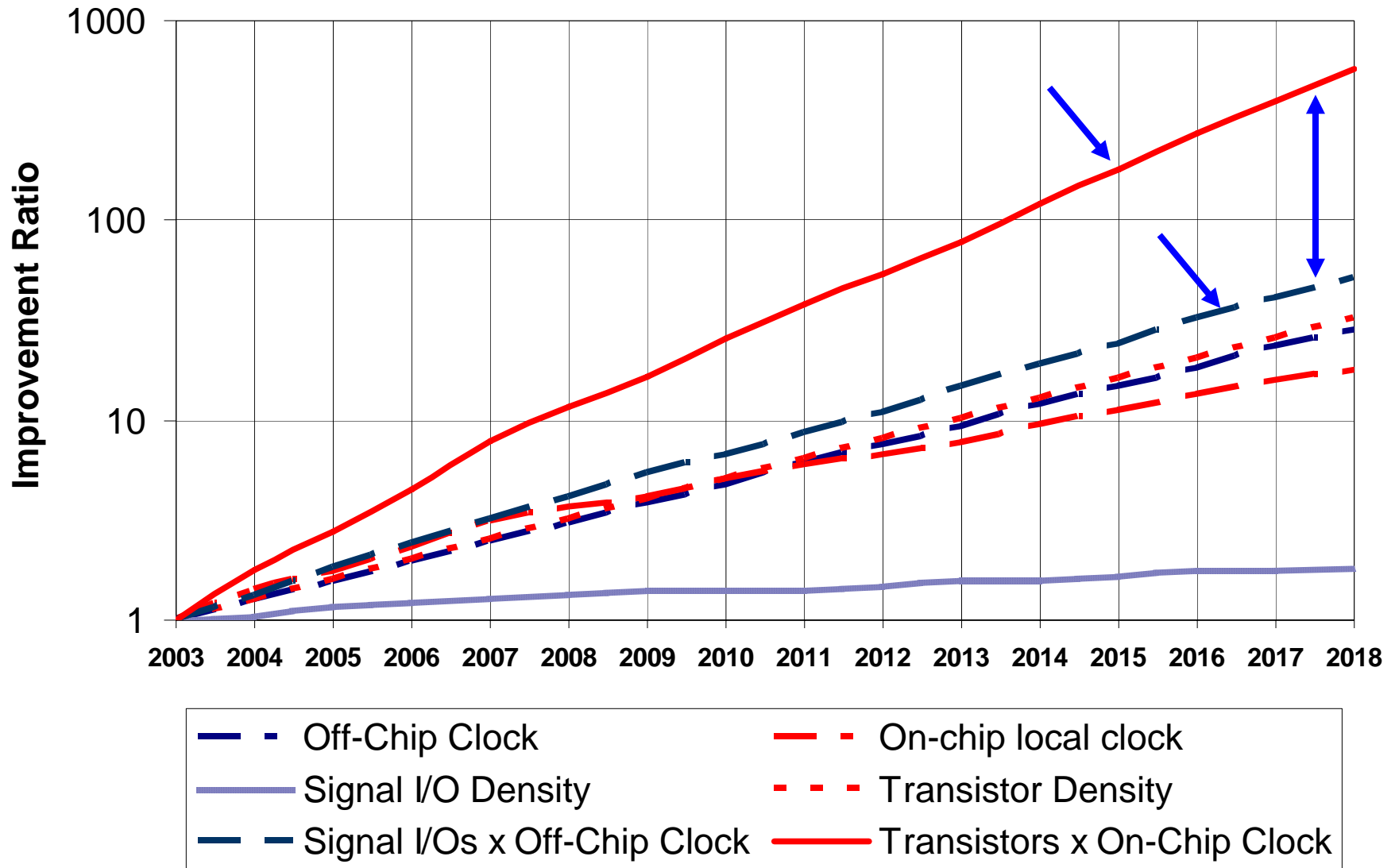
# Knee Curves with Today's Overheads



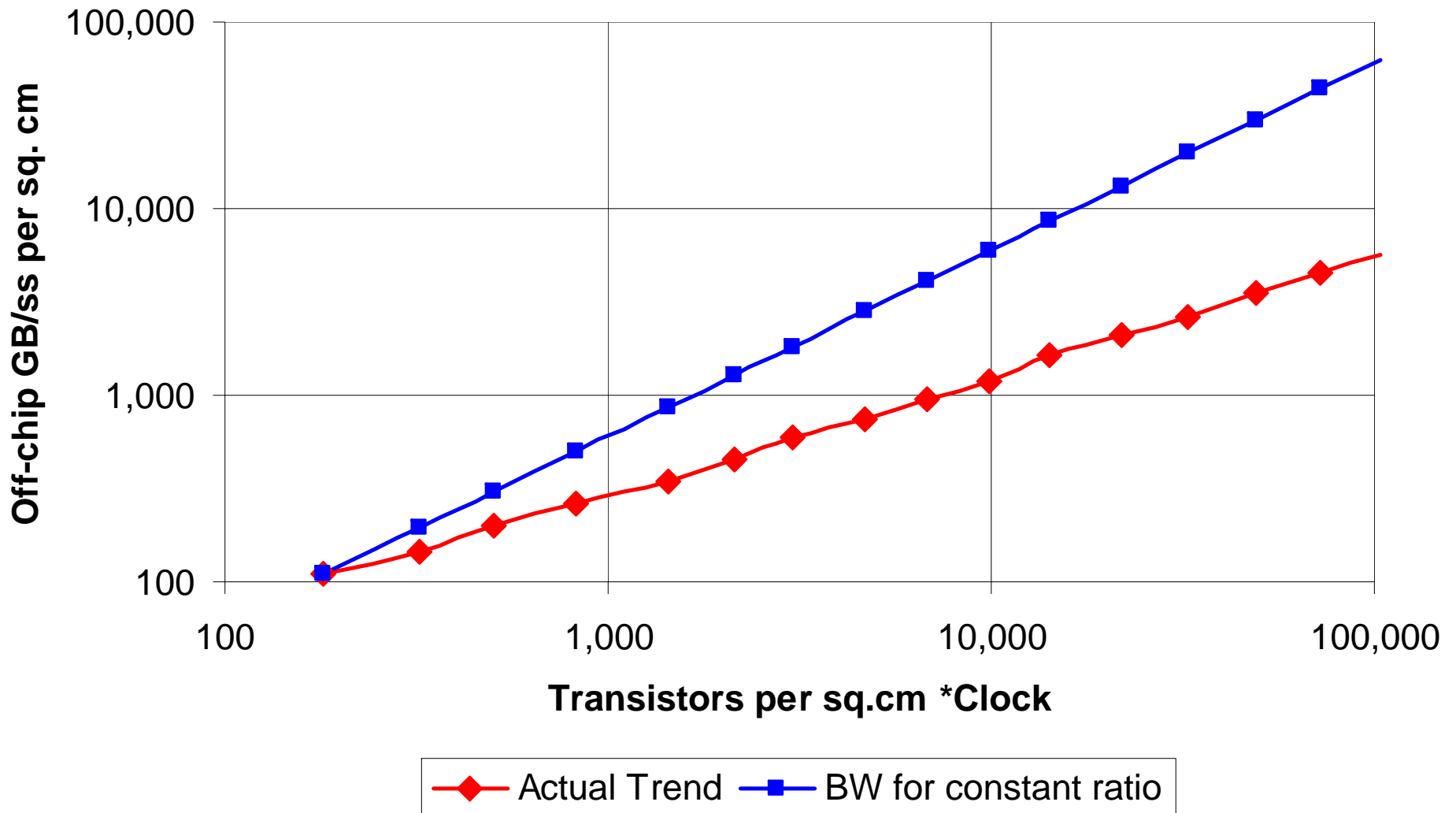
# Off Chip Bandwidth

- **Another measure: interchip bandwidth**
- **Upper limit = product of:**
  - # of off chip pins
  - Times % of pins not used as power/ground
  - Times max signalling rate per pin
- **Both # of pins/sq. cm and signal rate improve with time**
  - With 50% power/ground
  - But they don't match the growth in performance potential

# Relative Growth Rates



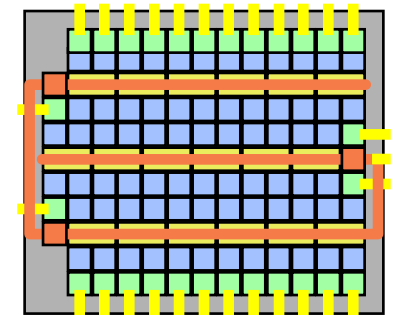
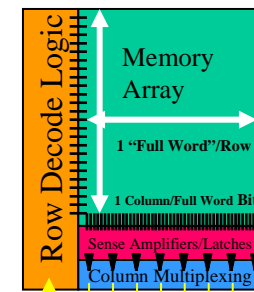
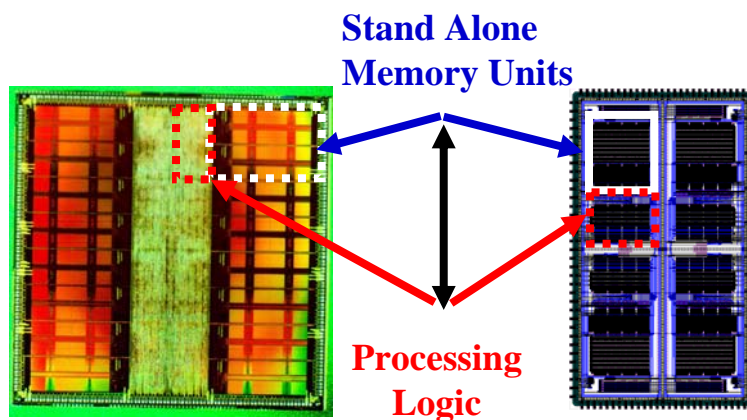
# Performance vs I/O



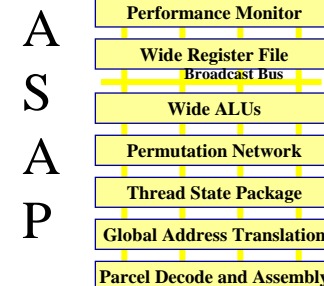
# Chip Level Architectural Design Space

# “Processing-In-Memory”

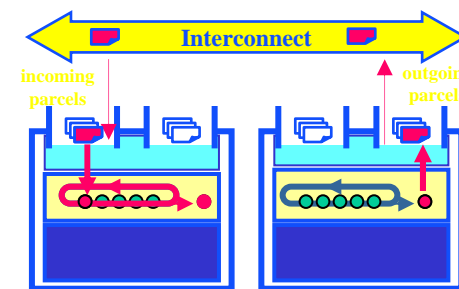
- High density memory on same chip with reasonable dense logic
- Very fast access from logic to memory
- Very high bandwidth
- ISA/microarchitecture designed to utilize high bandwidth
- Tile with “memory+logic” nodes



Tiling a Chip

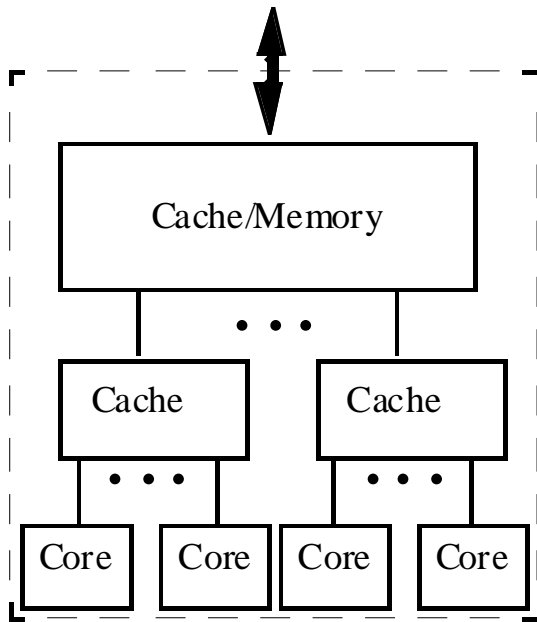


A Memory/Logic Node

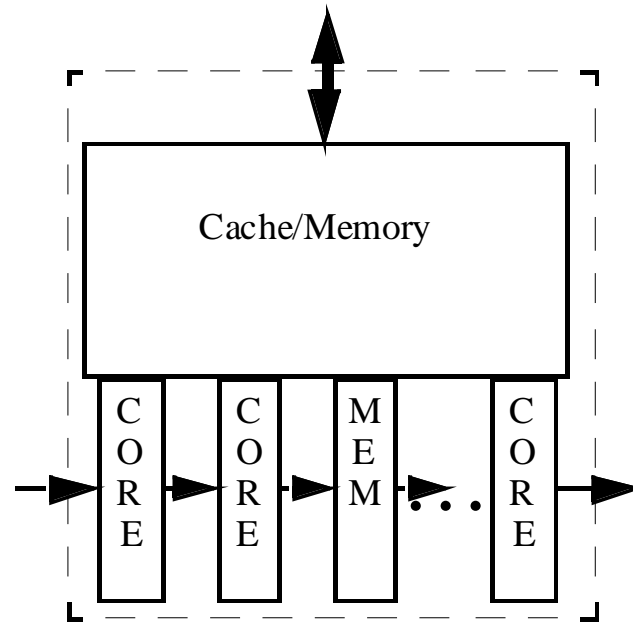


Parcel = Object Address + Method\_name + Parameters

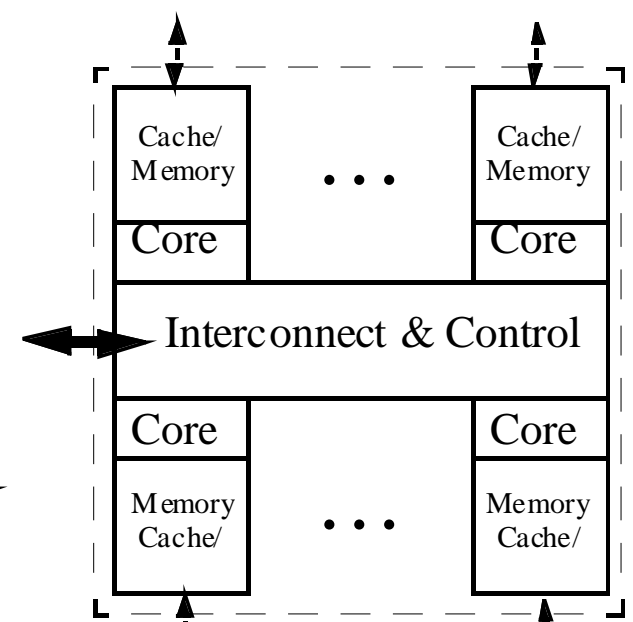
# How Might We Mix Logic & Memory



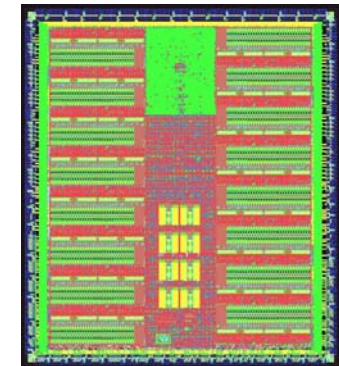
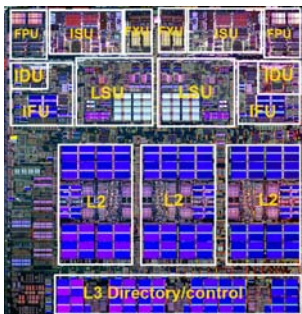
(a) Hierarchical Designs



(b) Pipelined Designs



(c) Array Designs



# Sample Chips

	<b>POWER5</b>	<b>X10q</b>	<b>Yukon</b>
<b>Year</b>	2003	2003	2002
<b>Technology</b>	0.13 Logic	0.13 Logic	0.15 DRAM
<b>0.18 Logic</b>			
<b>Area</b>	389mm <sup>2</sup>	??	??
<b>Type</b>	Hierarchical	Pipelined	Array
<b>Transistors</b>	276M	114M/62L	??
<b>Cores</b>	2@19% each	200=68%	256=14%
<b>Arch</b>	MT-SMP	Systolic	2D SIMD
<b>Core Clock</b>	2GHz	200 MHz	200MHz
<b>L2/Memory</b>	1.9MB=27%	23%	16MB=41%
<b>Contacts</b>	5,400	1,280	
<b>Memory</b>	41%	23%	44%
<b>Signal I/Os</b>	2,313	845	
<b># Ports</b>		10	
<b>Data B/W</b>	16GB/s	40Gbps	200MB/s
<b>Internal BW</b>			25.6GB

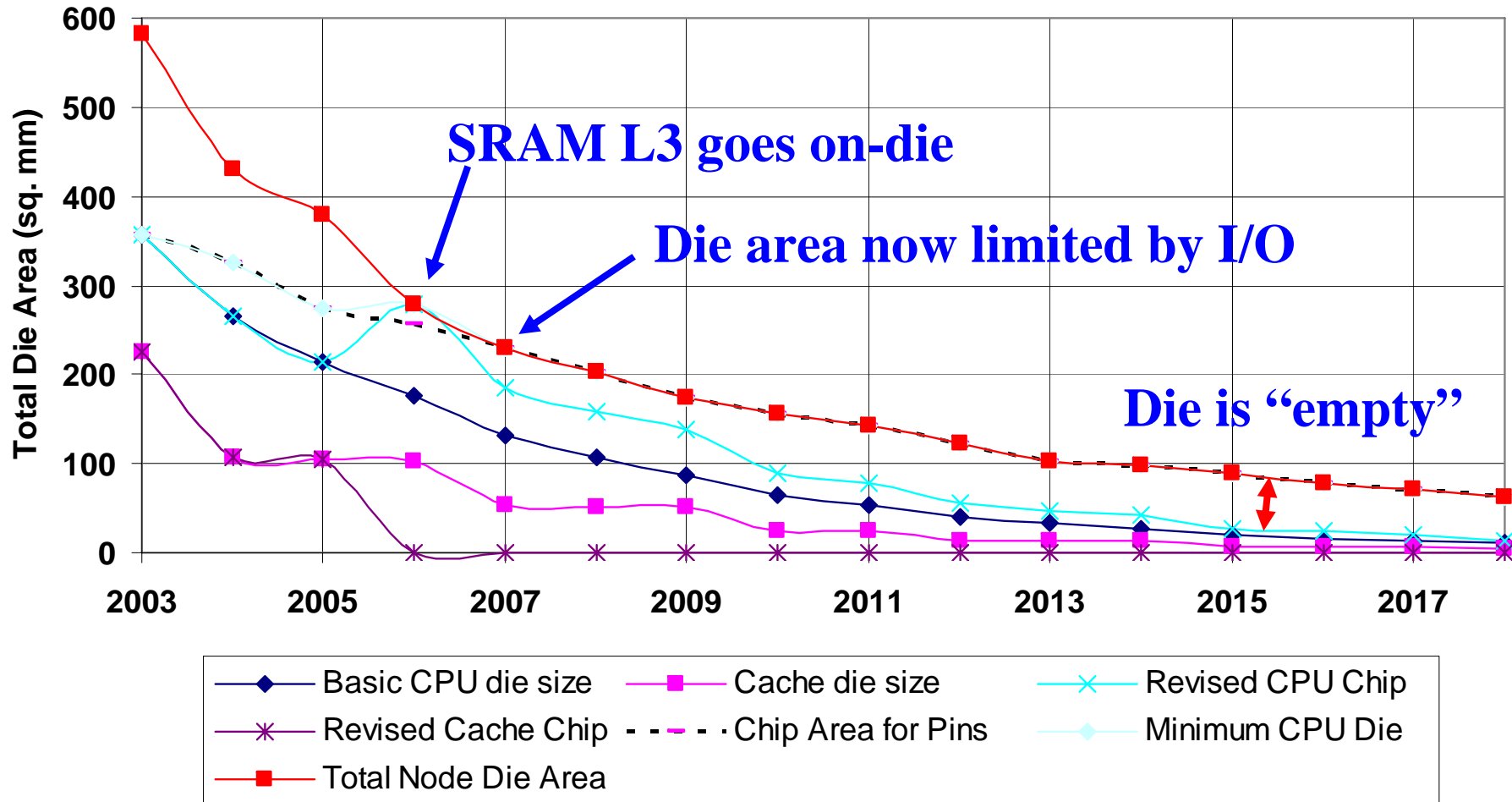
# Explorations

**Note: we only look at Hierarchical**

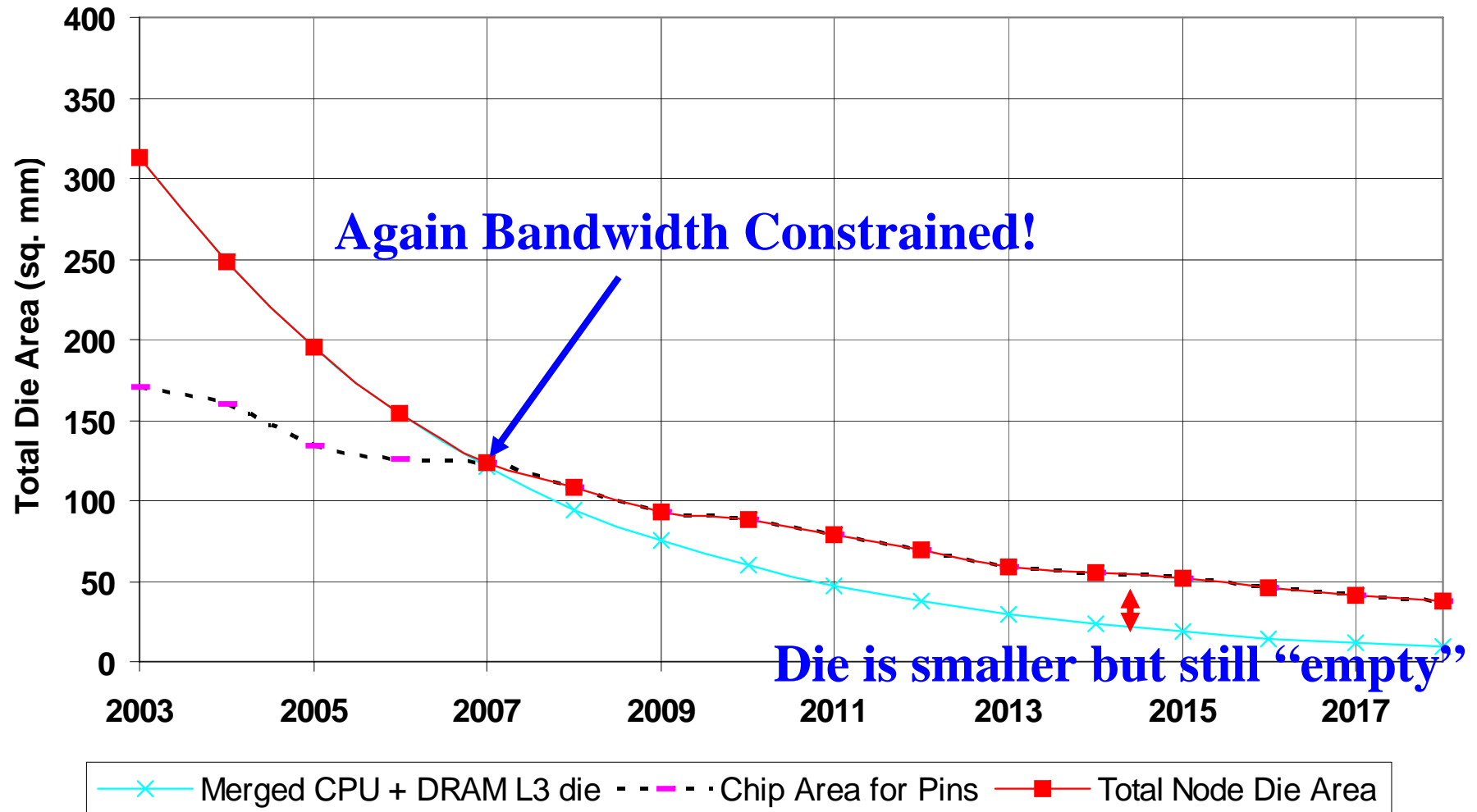
# Scaling Models to Explore

- **Shrink:** Take today & just shrink
- **Shrink & Merge:** replace L2/L3 SRAM with DRAM (& reduce clock)
  - Alternative would be EDRAM
- **Constant die size:** Add cores to fill die
- **Single chip type:** merge with memory
  - Ensure desired memory/performance ratio
- **Consider for each model:**
  - How many pins needed for constant bandwidth ratio

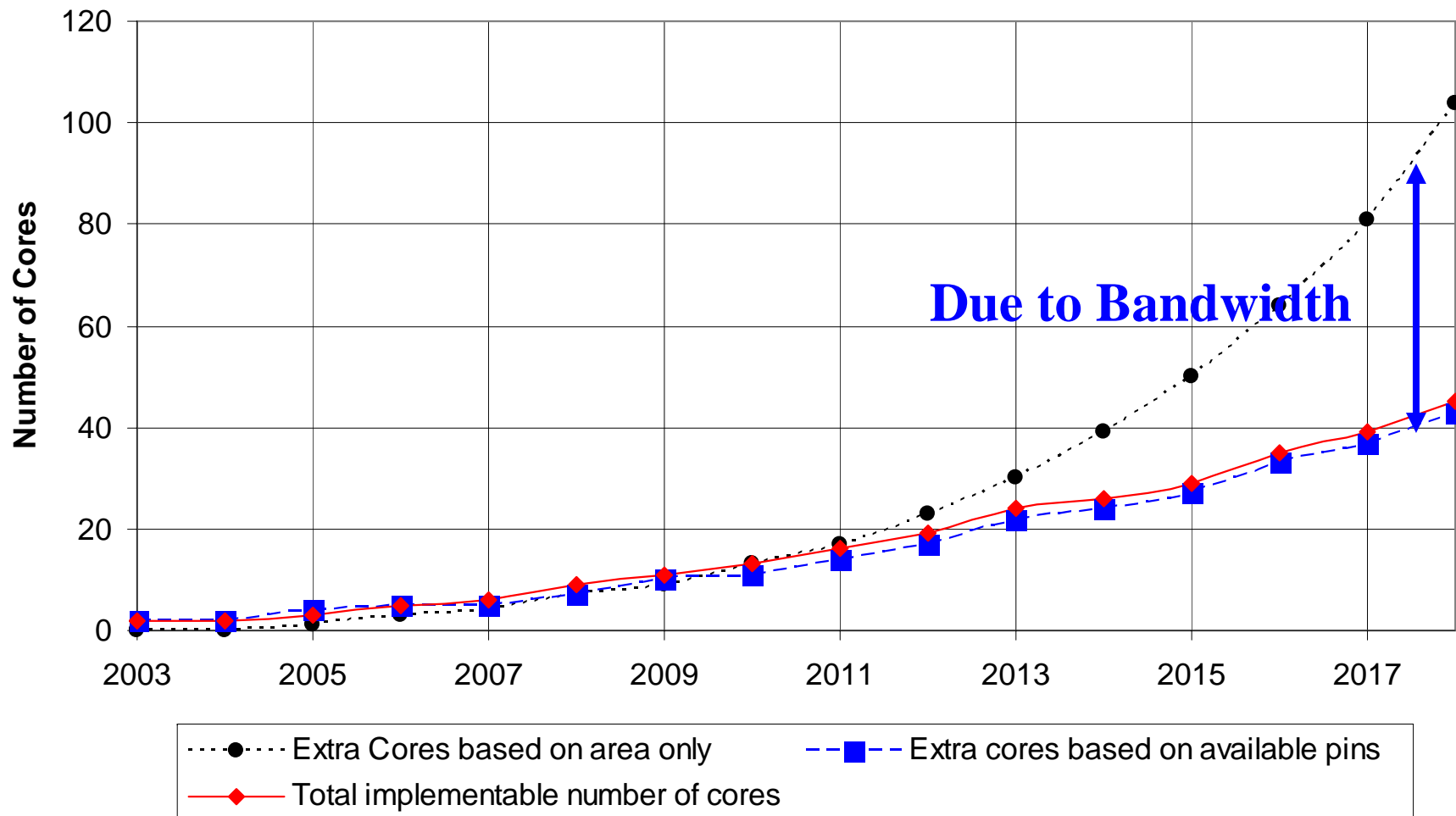
# Shrink Model



# Shrink & Merge Model

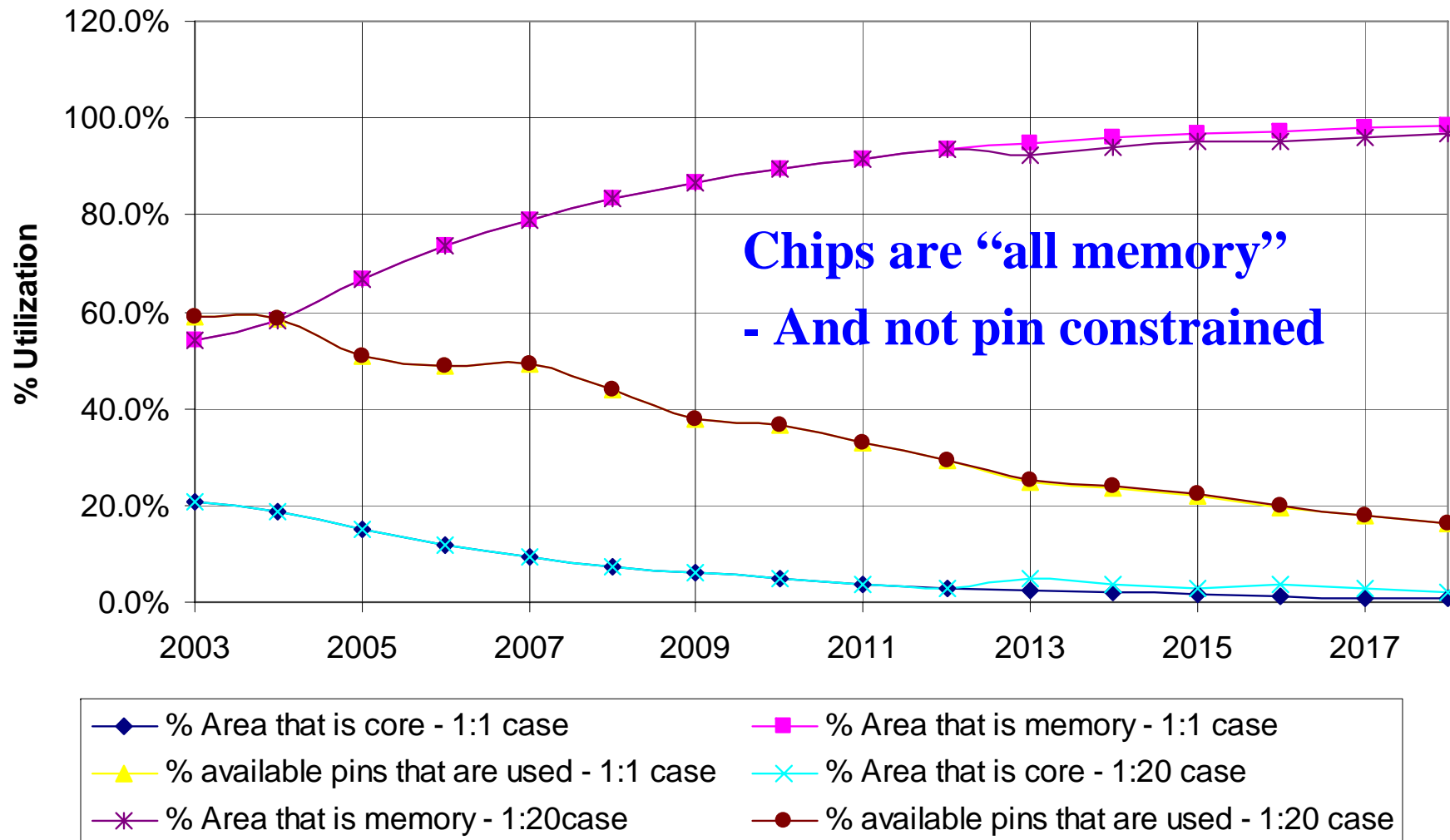


# Constant Die Model



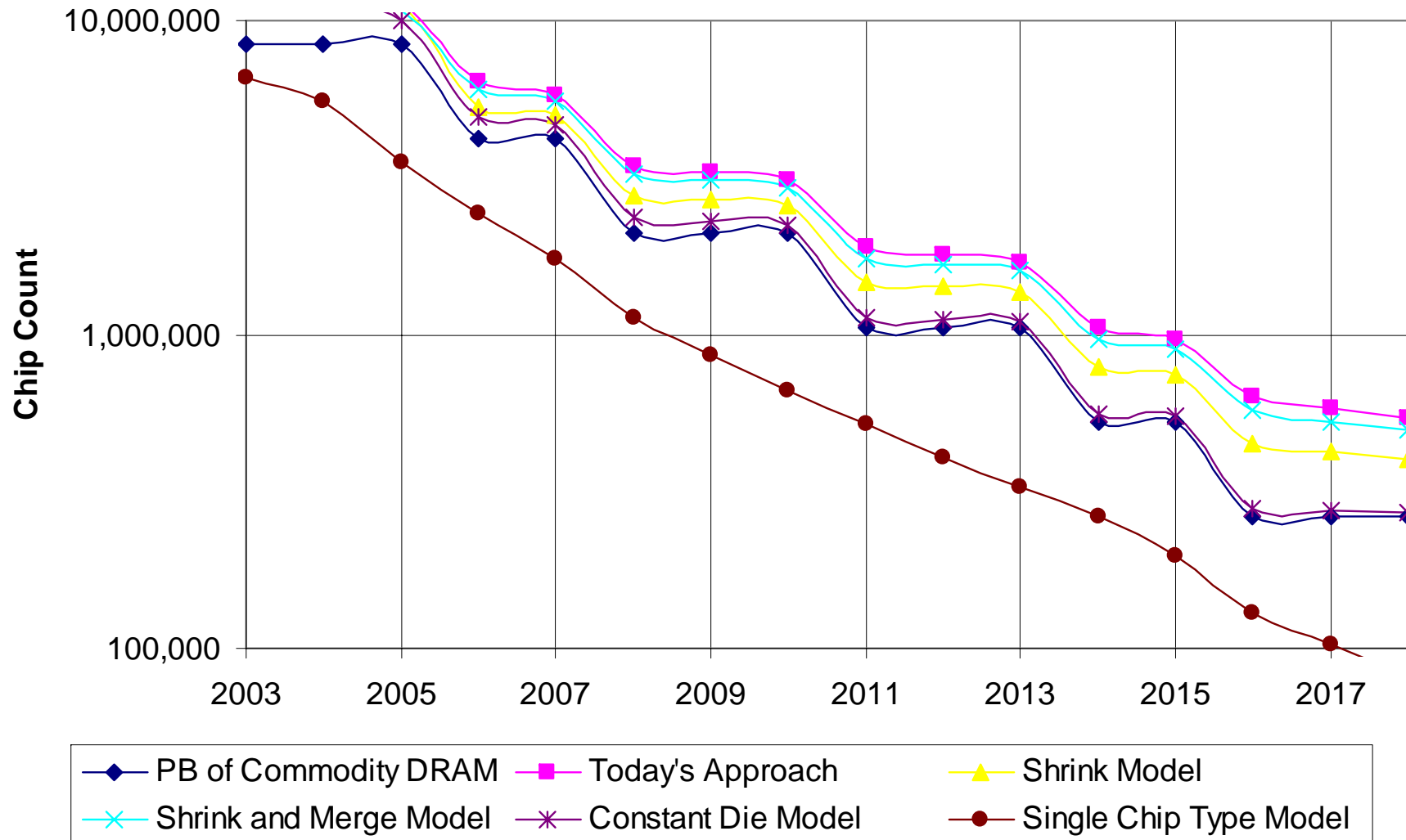
# Single Chip Type Model

(With Constant Die Size, Parameterized Bytes/Flops)

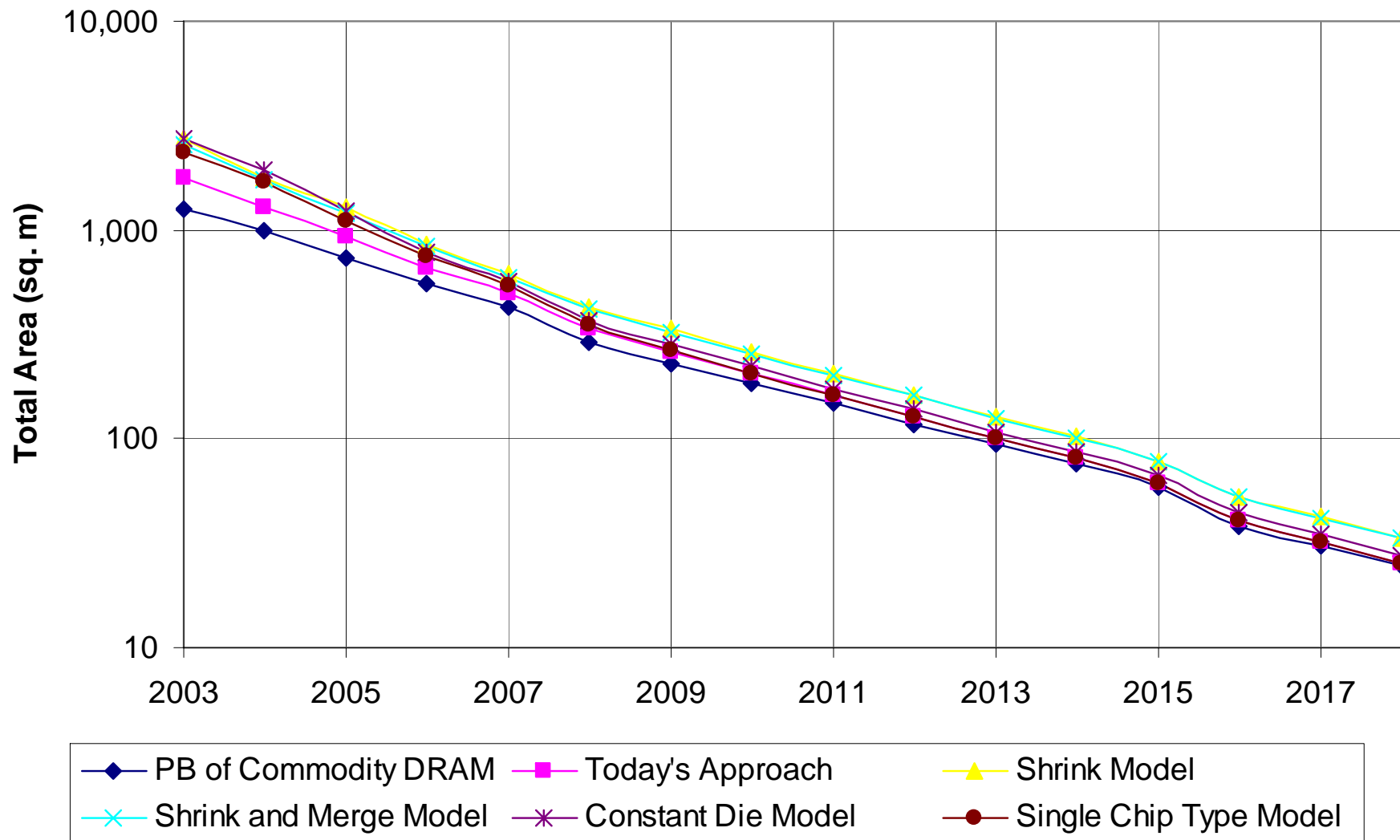


# Summary

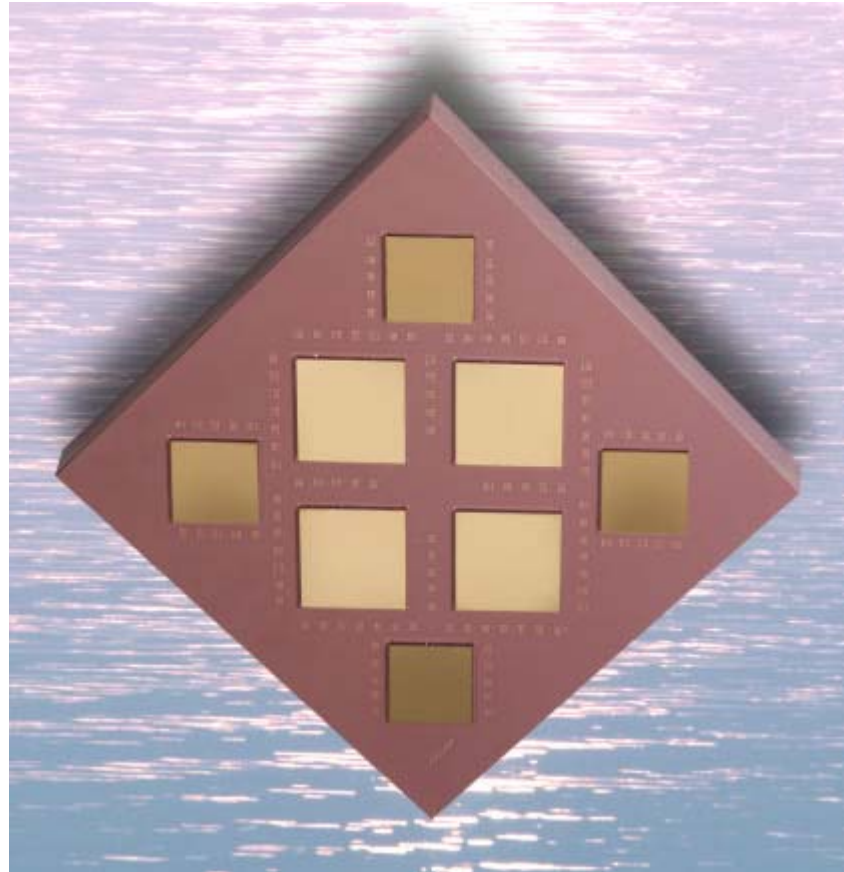
# Chip Count for a Petabyte System



# Silicon Area for a Petabyte System



# Silicon Alone is not the Complete Story



- Only 20% of MCM is silicon
- And we haven't accounted for the heat sink!

# Summary

- **Prior analysis still superficial**
  - **Need to include power limits**
  - **Degrade real performance growth rates**
  - **Account for all fab differences**
  - **Repeat for other CMP models**
- **But clear conclusions**
  - **The bulk of silicon by area = memory**
  - **Thus use densest available memory**
  - **Off-chip bandwidth the next constraint**
  - **Aim towards single part type die**
  - **Grossly overprovisioning processors is cheap by area**

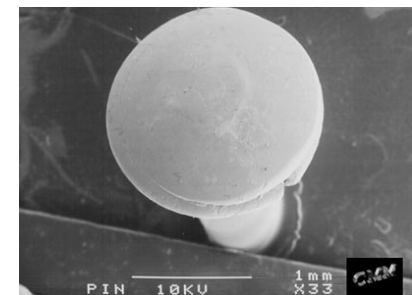
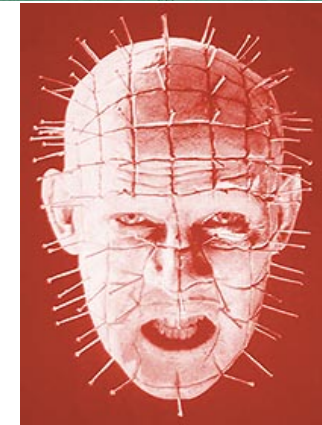
# How Do We make It Better?

- **Focus on “cheap” logic in dense memory fab process**
  - Don’t fret the clock rate
- **Reduce thread state**
  - Cost of moving/copying state = line reference
- **Simplify cores and “overprovision”**
  - “Pitch-match” to memory macro
- **Use on-chip memory for Data Recorder**
- **Change execution model from “named” core to anonymous core “nearest” memory object**
  - A “Traveling Thread” need never “wait” for processing resources

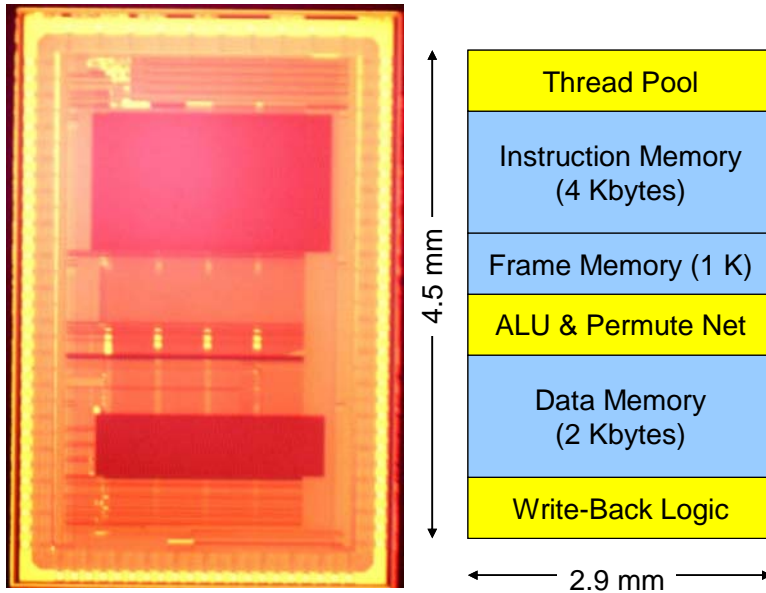
# The Original Question:

## How many Cores can Fit on the Head of A Pin?

- Area of a pin = .015 sq. cm.
- Assume Darkhorse 8051 @ 7 KT
- 2018: 4200 cores, @ 53 GHz
  - = approx 20 TOPS
- But to make them dance we need memory
- At 50/50 Memory & Logic
  - 2100 Cores + 100MB
- New Term: ***PIMHEAD***



# An Alternative Core: PIM Lite



- “Looks like memory” at Interfaces
- **ISA: light weight multithreaded+SIMD**
  - “Thread” = IP/FP pair
  - “Registers” = wide words in frames
- **Designed for multiple nodes per chip**
- **Permits “traveling threads”**
- **1 node logic area ~ 10.3 KB SRAM**
- **TSMC 0.18u 1-node 1<sup>st</sup> pass success**
- **3.2 million transistors (4-node)**

