The Driving Force for ICT Development in this Decade: Power Consumption

GreenTouch Open Forum

Seattle, Washington November 17, 2011

History

Power has always been important for Electronics

- The initial computers were built with vacuum tubes and power requirements placed an upper limit on size and performance.
- The introduction of the transistor dropped power dramatically and enabled a new generation of products.
- The integrated circuit brought a 3rd wave of power reduction enabling the introduction of more powerful portable electronics.
- We are now entering a 4th wave of innovation with the introduction of 3D ICs and heterogeneous integration.

The 45 year history of knowing what comes next is over

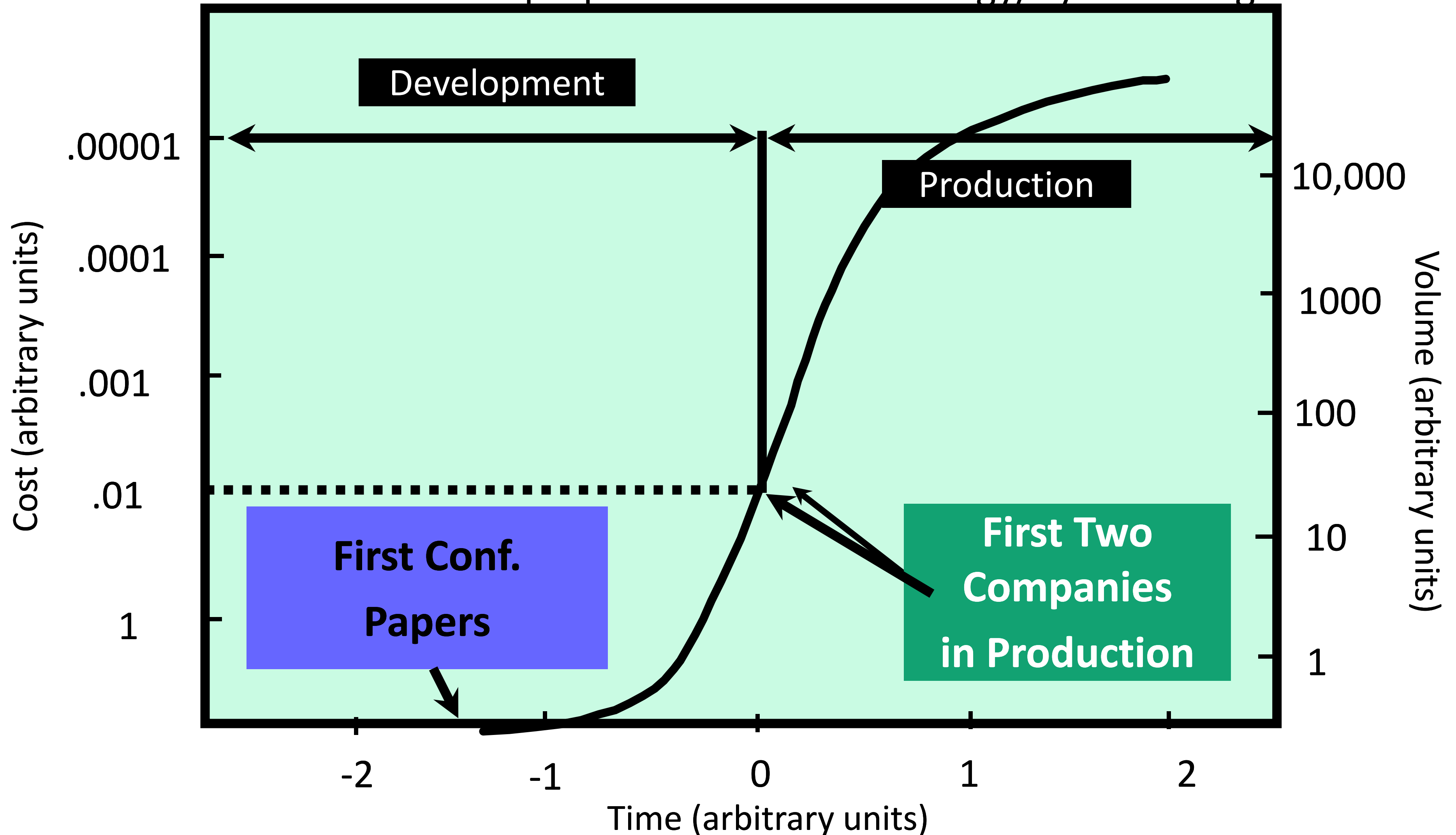
Progress has been paced by Moore's Law and driven by:

- **Focus was on design and fab**
  - **Shrinking geometries**
  - **Expanding wafer size**
  - **Higher density designs**

For digital circuits there are now limitations that can't be met by these activities alone.

Semiconductor Electronics has been Characterized by an S-Curve

Production Ramp-up Model and Technology/Cycle Timing



Semiconductor Electronics has been Characterized by an S-Curve

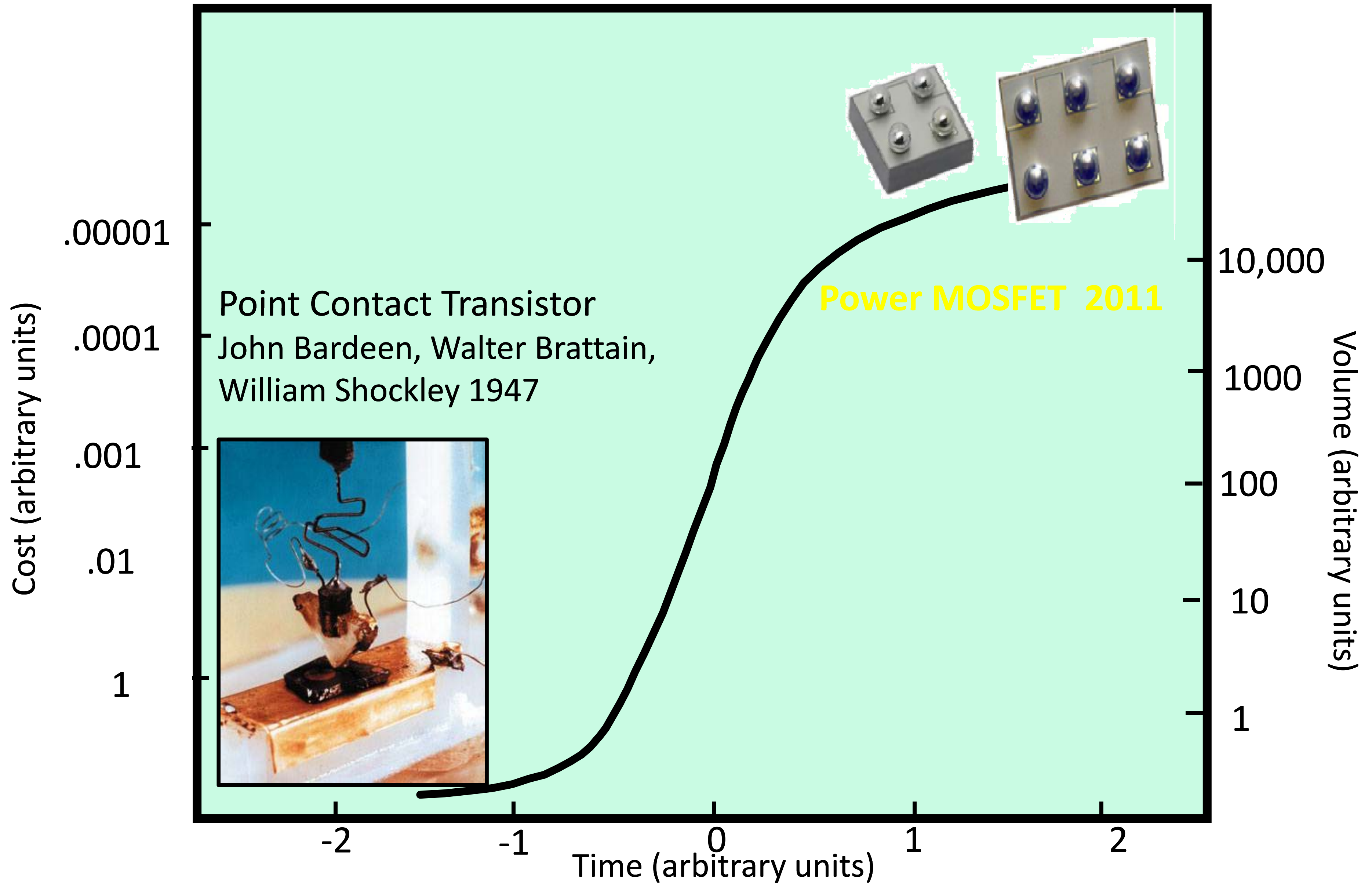
Production Ramp-up Model and Technology/Cycle Timing

In the first S-Curve cycle in the semiconductor industry the Technology was the transistor

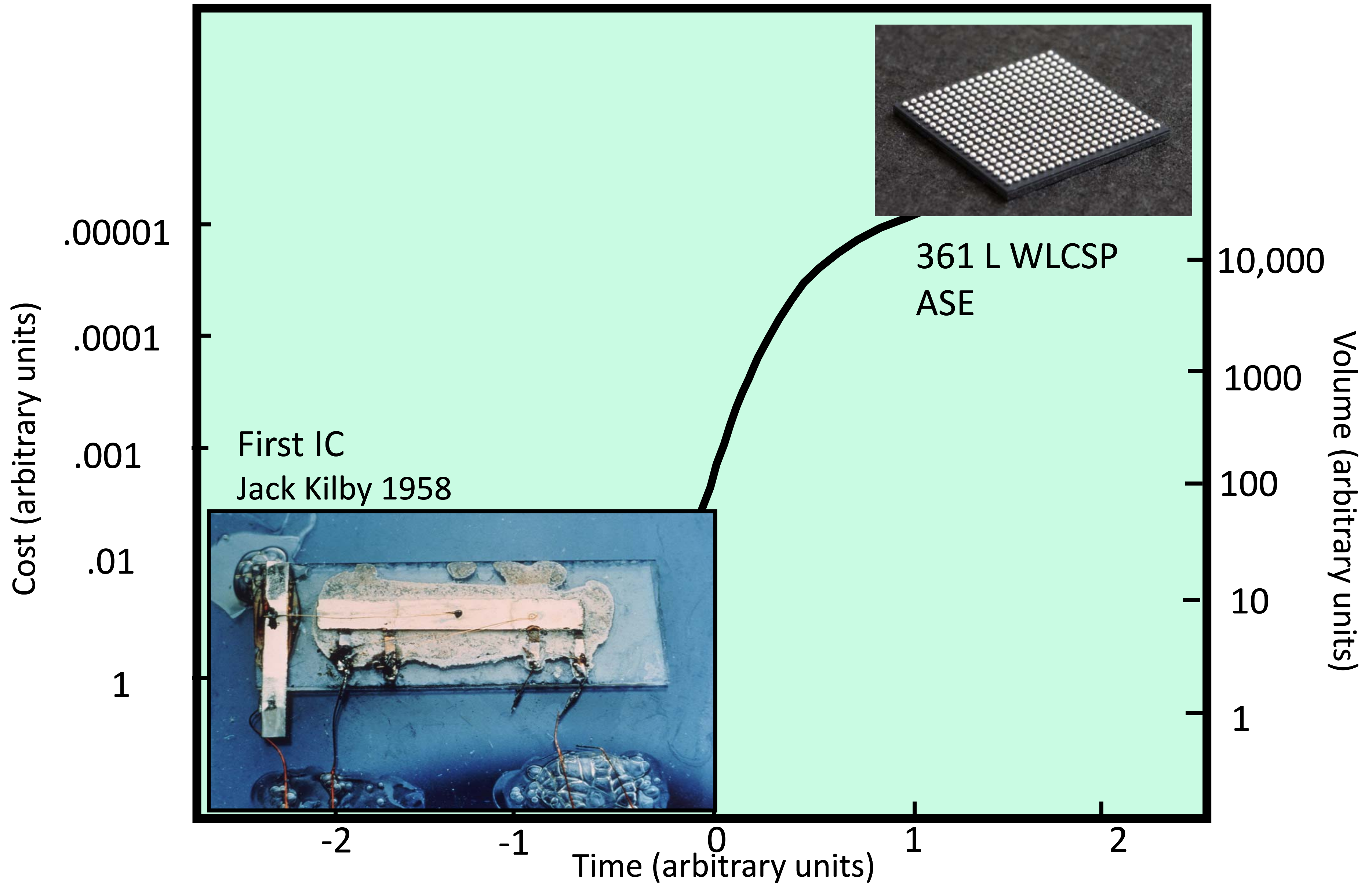
-2 -1 0 1 2

Time (arbitrary units)

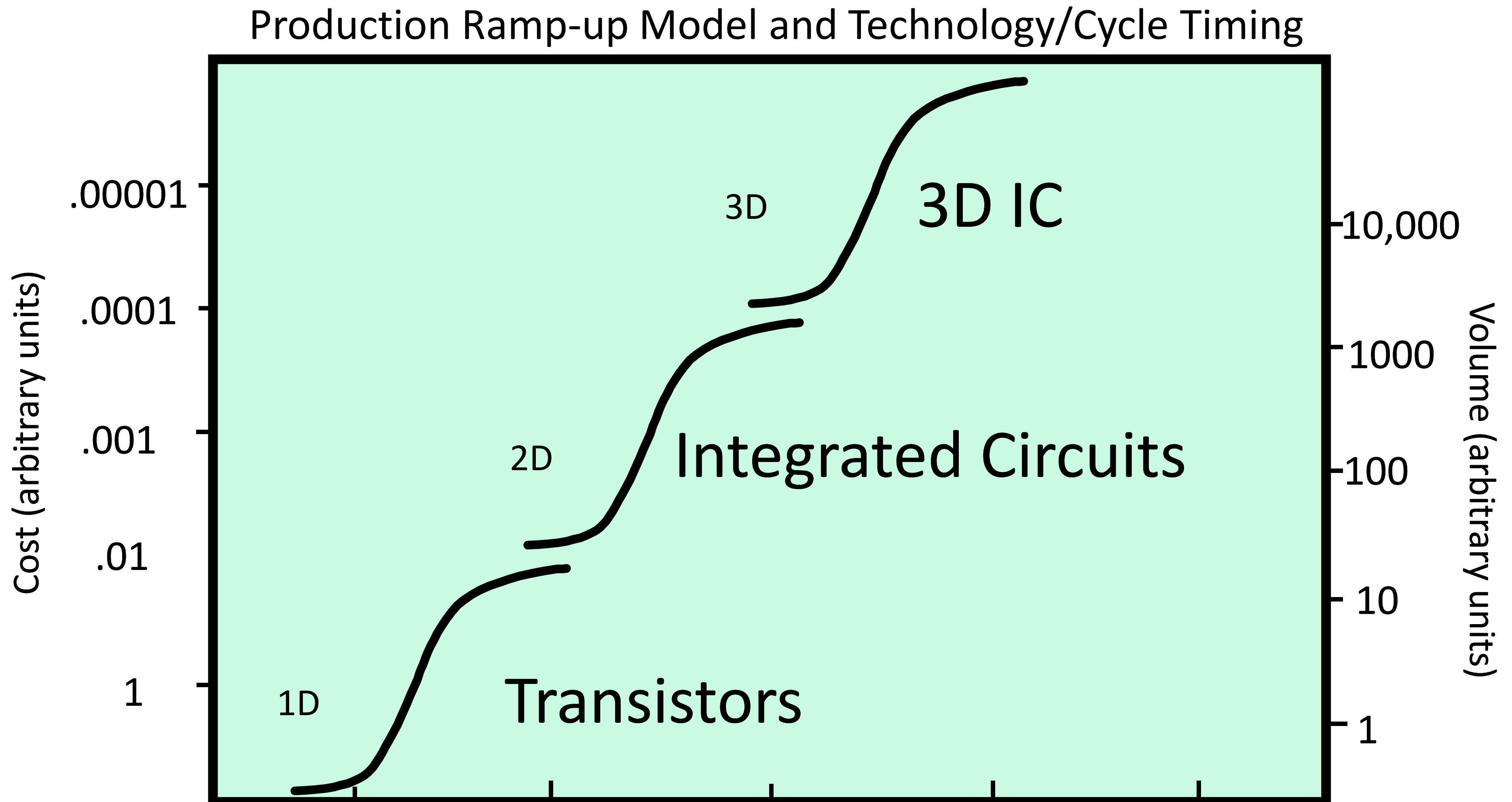
The Transistor S-Curve



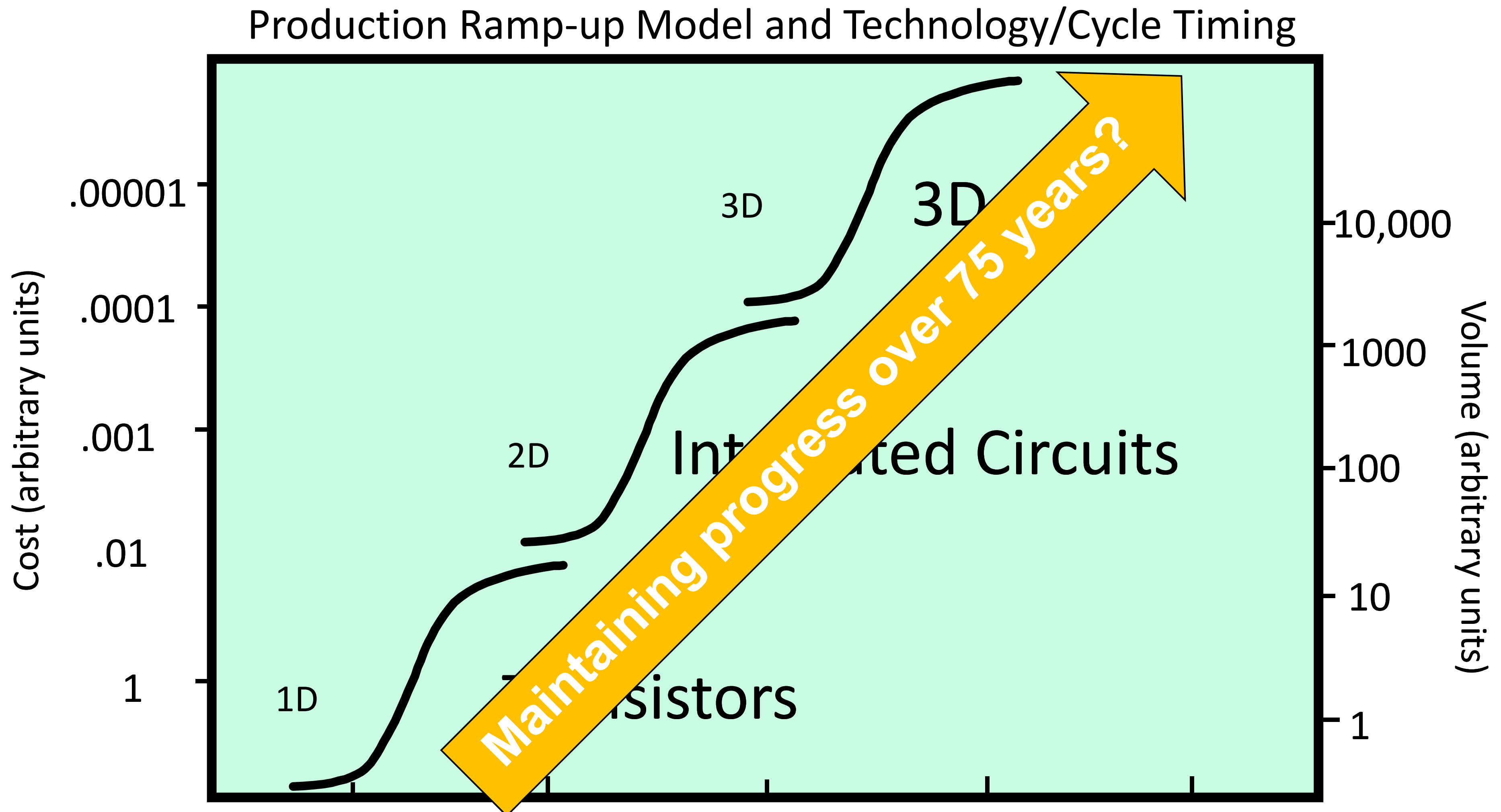
The Integrated Circuit S-Curve



Can the 3D IC maintain Progress through a third S-Curve Cycle?



Can the 3D IC maintain Progress through a third S-Curve Cycle?



Reduced Benefit from Shrinking CMOS

- Gate tunneling current increases
- Subthreshold channel leakage current increases
- Device parameter variability increases
- Source/drain resistance increases
- Copper interconnect resistivity increases

Power no longer scales with feature size, both static and dynamic power dissipation increase due to these barriers.

Consumer Market Demands now drive Power Requirements

- Power reduction initially driven by battery life and size/weight of portable electronics
  - Digital cameras, cell phones, walkman, etc.
- There has been remarkable growth in consumer driven demand for data as smart phones, digital television, 3D television and streaming video emerged
  - 3D holographic telepresence is not far behind

The Major Challenges

- ✓ Bandwidth Density
- ✓ Power requirements
  (power density and power integrity)
- ✓ Latency
- ✓ Cost

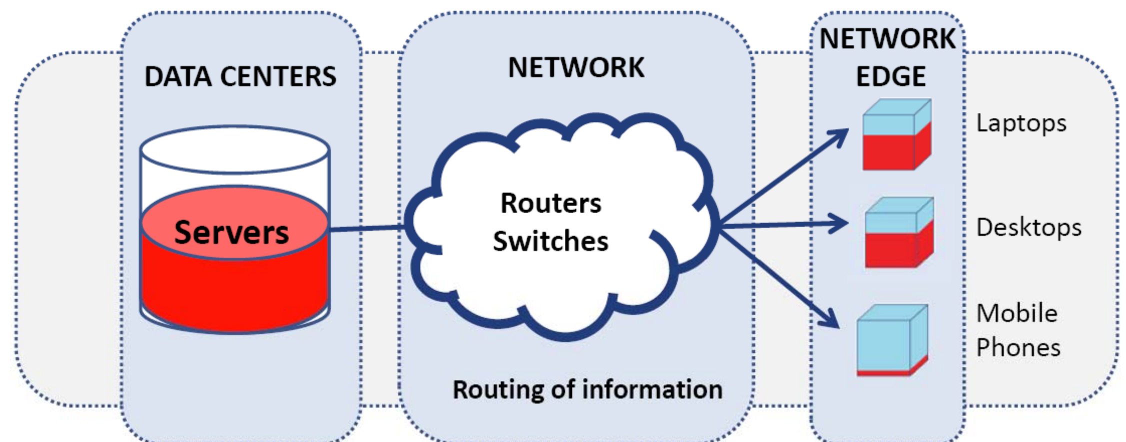
The Key to Power Optimization

Use power only where it is needed and when it is needed

In order to optimize we must identify each component of power use.

ITRS has a focus on the electrical components; INEMI has a focus on the box level network components. If the power use by each component is known, these organizations can focus their activity on identification of potential solutions to reduce power where it is needed most:

- **Memory**
- **Processors**
- **Serdes**
- **E to O to E conversion**
- **Fiber transmission losses**
- **Data Centers**
- **Other**



The Key to Power Optimization

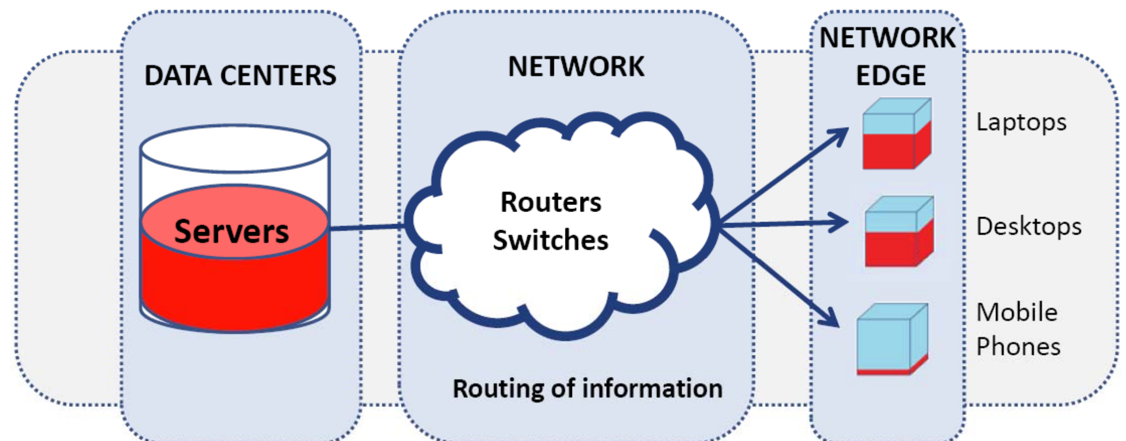
Use power only where it is needed and when it is needed

In

IT k
le
th
so

ITRS and INEMI will not address the Topology and architecture of Networks but the component work may enable new solutions in network architecture to reduce power.

- Memory
- Processors
- Serdes
- E to O to E conversion
- Fiber transmission losses
- Data Centers
- Other



How can we reduce power

- Optimize network topology/architectures
- Operate the system with DC power
- Reduce photonics power requirement
- Move photonics closer to the transistors
- Reduce semiconductor power requirement

Today the ITRS addresses semiconductor devices and packaging in detail and photonics at the packaging level only.

Reduce semiconductor power

- Reduce leakage currents (new transistor designs)
  - Transistors are less than 10% of IC power today and going down
- Reduce on-chip Interconnect power by:
  - Improved conductor conductivity (new material)
  - Decrease capacitance (new material)
- Reduce interconnect length (3D integration)
- Reduce operating frequency (increased parallelism)
- Reduce operating voltage (enabled by lower frequency)
- Voltage regulator per core (power when & where needed)
- Reduce high speed electrical signal length (serdes)

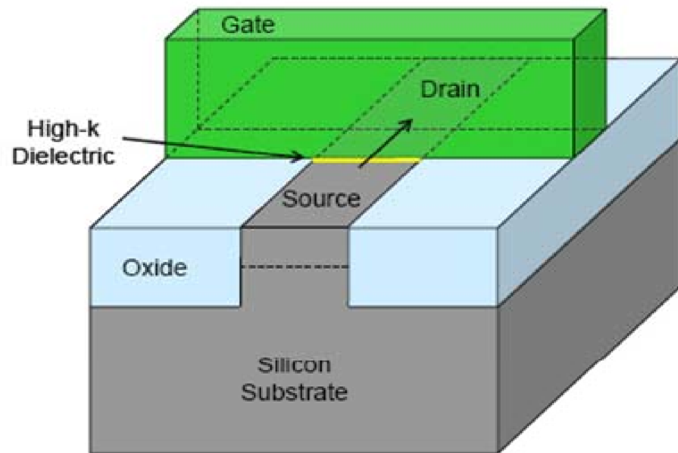
Reduce semiconductor p...

- Reduce leakage currents (...designs)
  - Transistors are less than 10... ...y and going down
- Reduce on-chip Inte... ...by:
  - Improved cond... (new material)
  - Decrease ca... (new material)
- Reduce i... ...th (3D integration)
- Redu... ...quency (increased parallelism)
- R... ...voltage (enabled by lower frequency)
- ...ator per core (power when & where needed)
- ...gh speed electrical signal length (serdes)

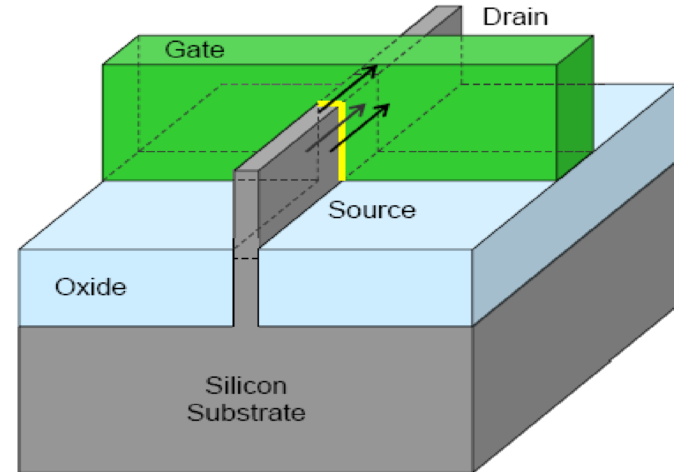
Reliability will be more difficult due to transistor wear out

Reduce Leakage Currents

- **High k/metal gates** (Thicker gate dielectrics)
- **Multi-gate designs**



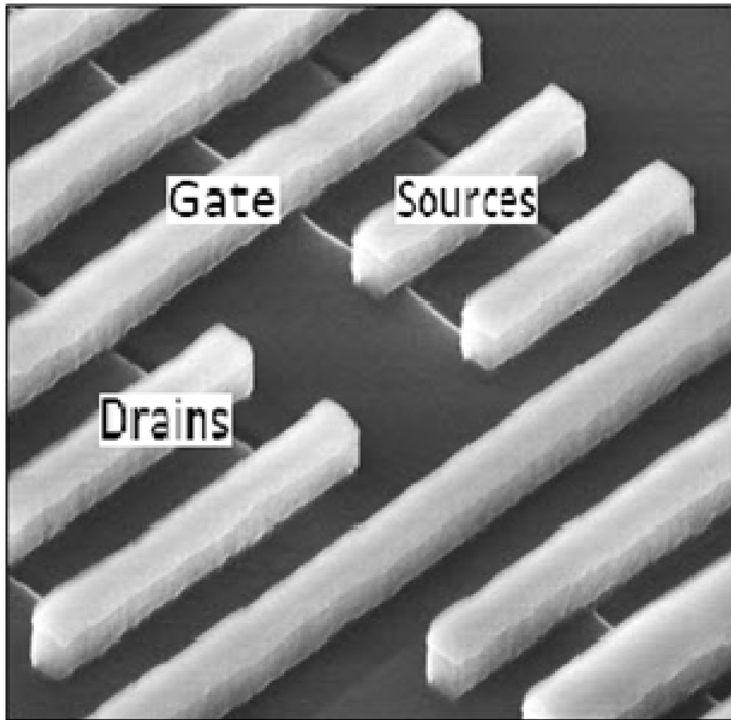
Traditional Transistor
High k/metal gate

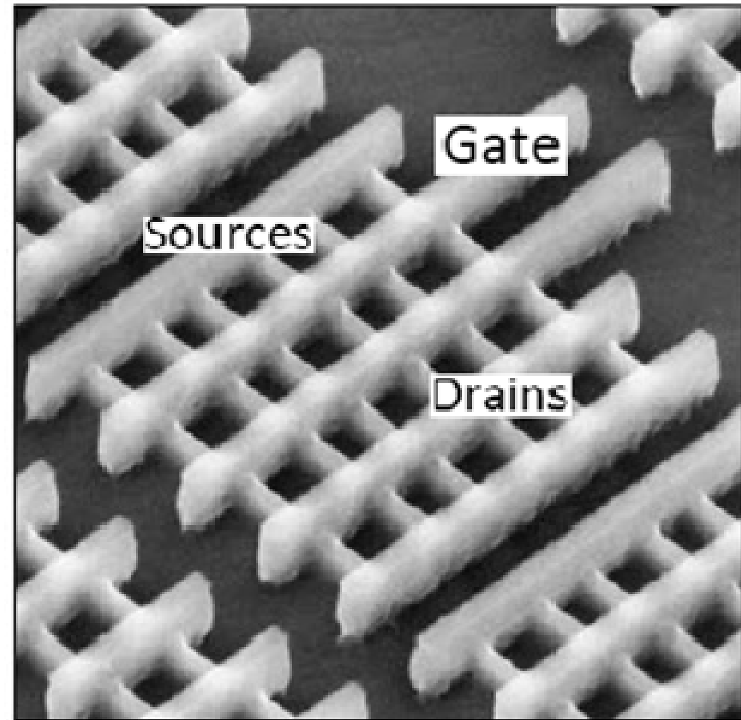


Triple-gate FinFET
High k/ metal gate

Reduce Leakage Currents

- High k/metal gates (Thicker gate dielectrics)





32nm
Planar Transistors

22nm
Triple-gate Transistors

Interconnect has become the limiting factor in power requirement

- **Moore's law scaling has decreased transistor delay and power by three orders of magnitude while interconnect delay and power dissipation have been negatively impacted.**

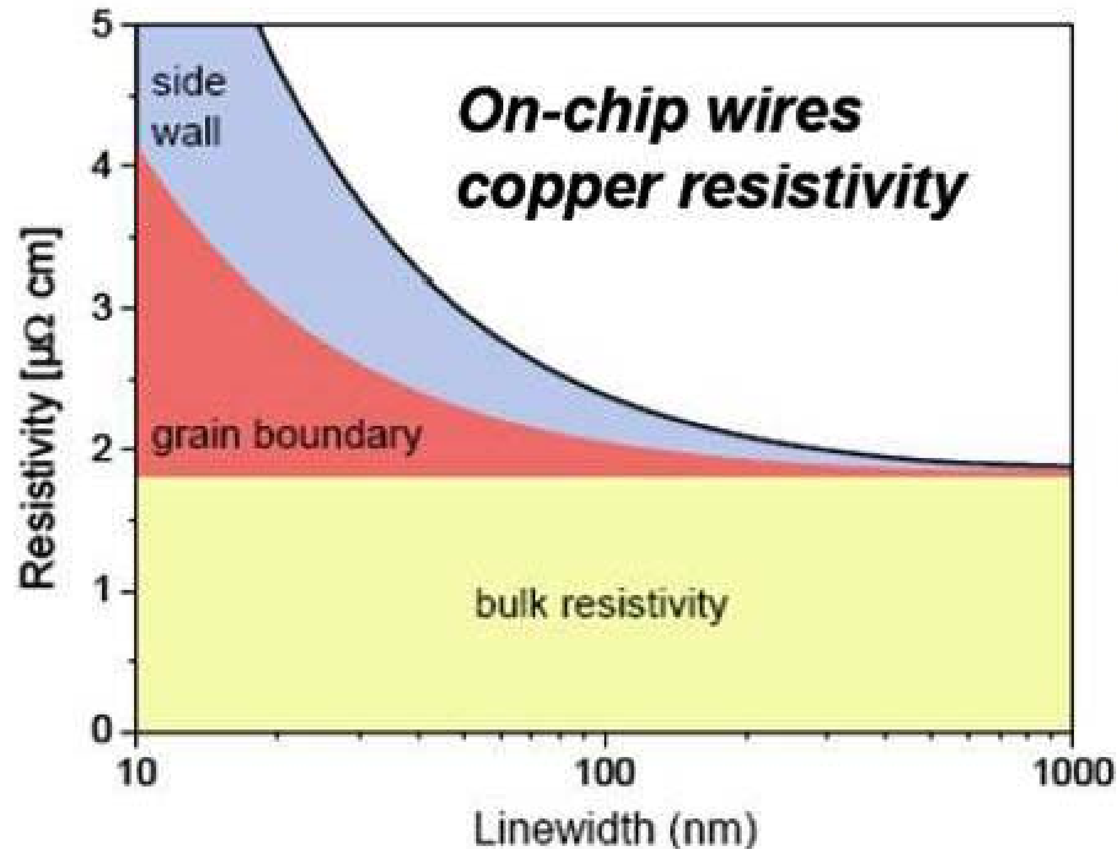
Change from Al/SiO_2 to Cu/Low k can solve the problem

For small dimensions metal resistivity increases rapidly due to sidewall and grain boundary scattering increasing RC delay

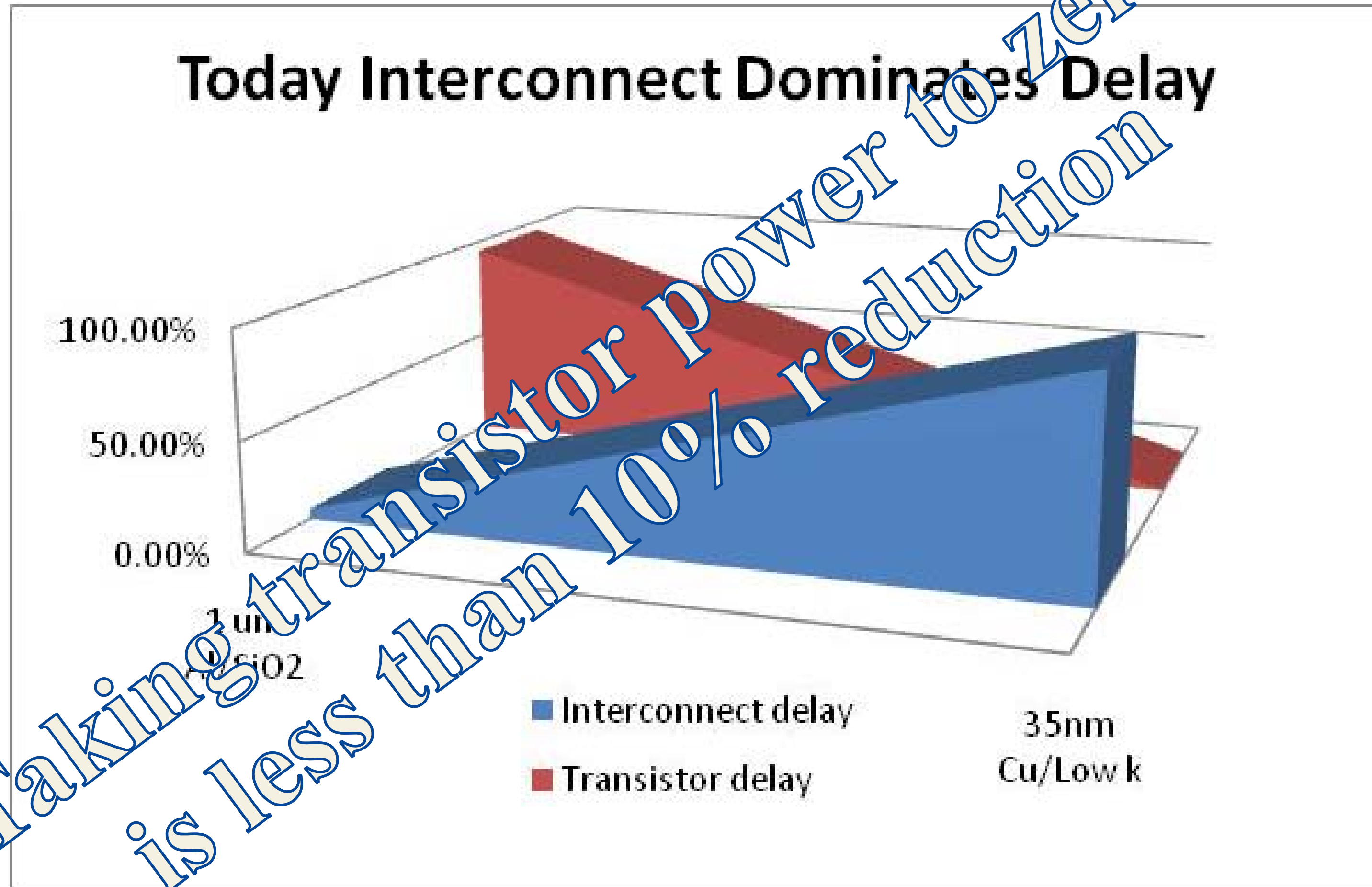
- For 1.0 um Al/SiO2 technology, transistor delay was 20 psec and RC delay for a 1 mm line was 1.0 psec
- For 32 nm Cu/low k technology, transistor delay will be 1.0 psec and RC delay for a 1 mm line is 250 psec

Copper Conductivity is reduced

- **Grain boundary and side wall scattering increase resistivity as linewidth decreases**

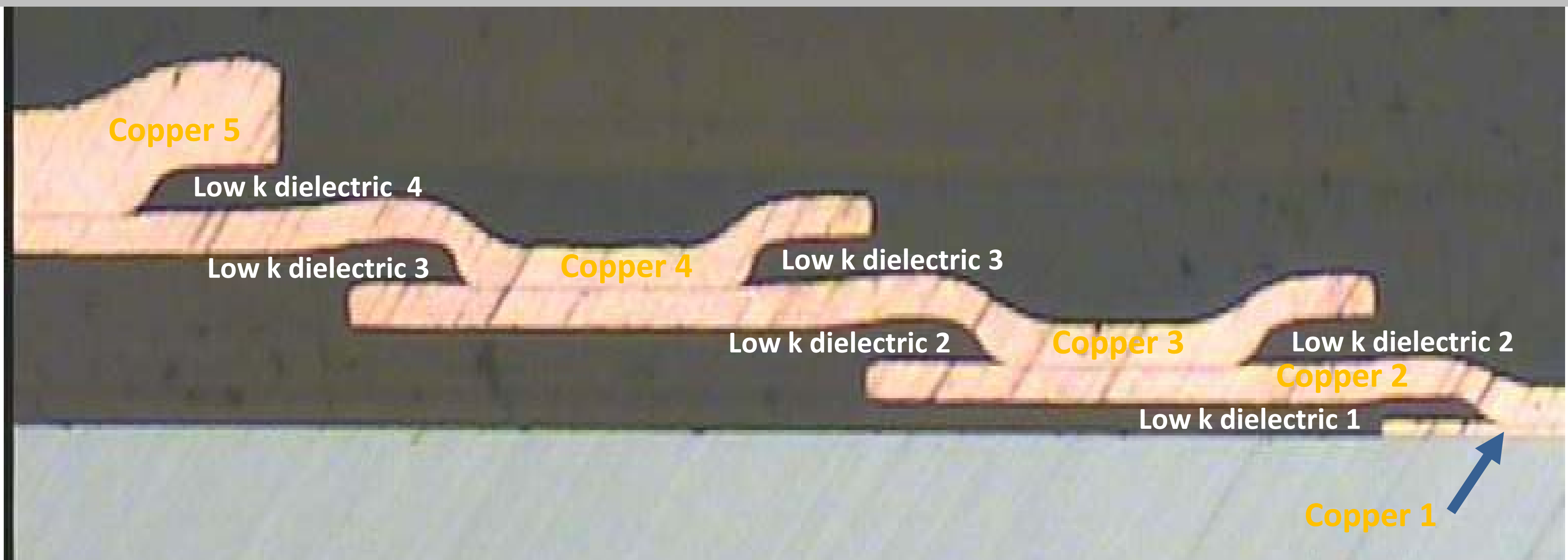


Interconnect will dominate delay and power dissipation



Ultra Low K Dielectrics for Wiring Layers

Spin-on ULK with k=1.8 is now in qualification in a high volume fab



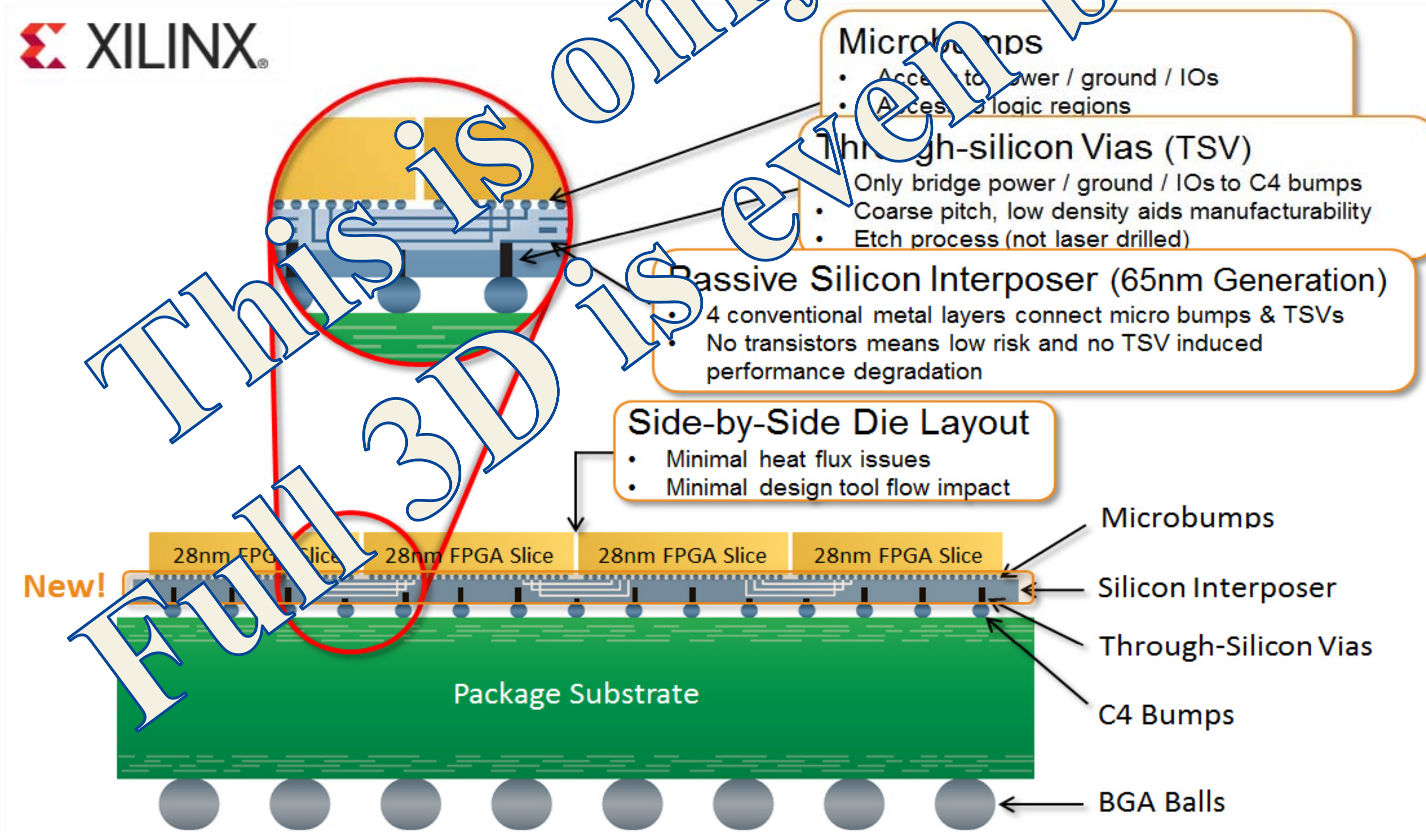
Reducing K from 3 to 2 provides a .33 reduction in Power

Example: Xilinx Stacked Si FPGA

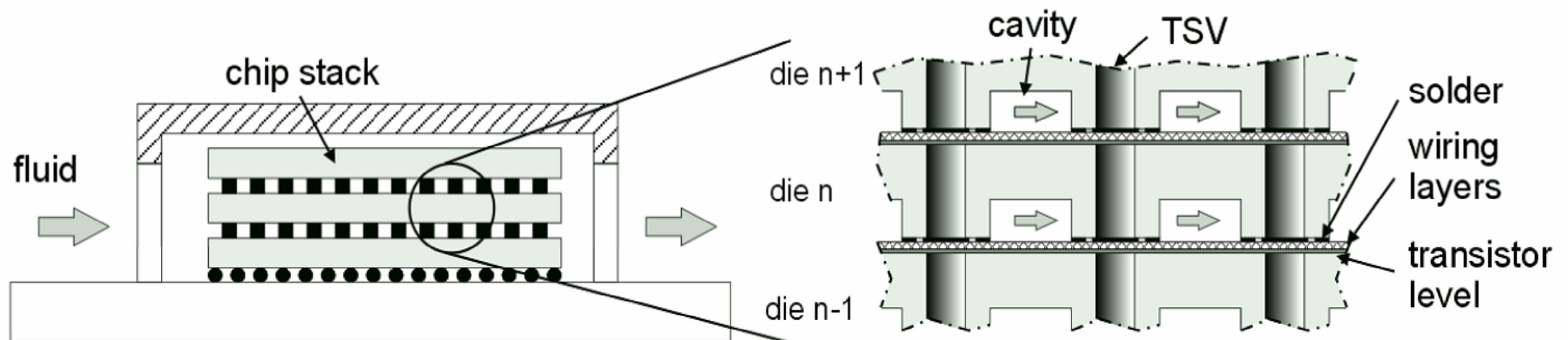
- Interposer substrate has more than 10,000 routing connections
- Compared with standard I/O connections it provides:
  - > 100X die-to-die bandwidth per watt
  - one-fifth the latency
  - Uses no high-speed serial or parallel I/O resources.

This is only 2.5D

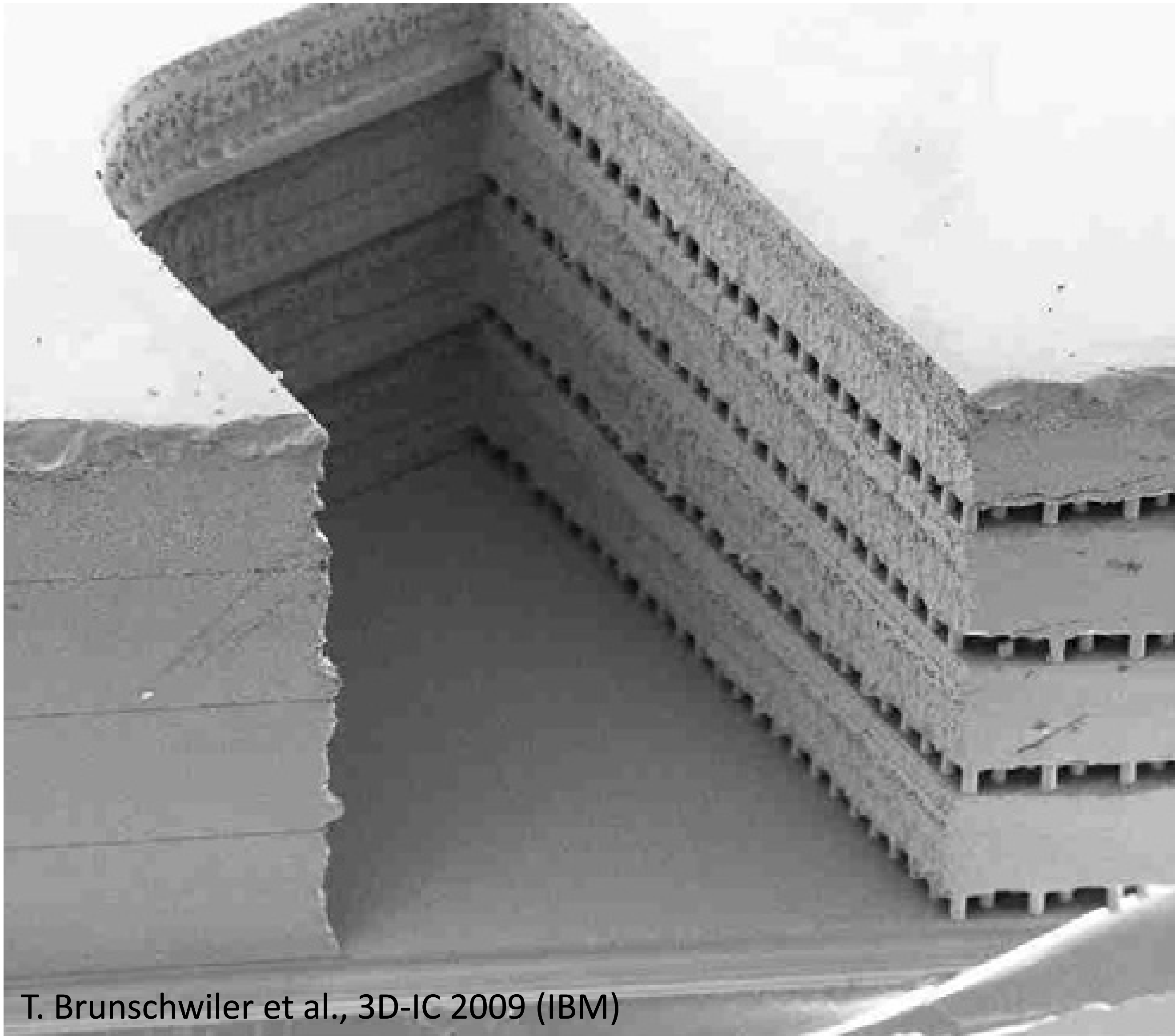
Full 3D is even better



We still must handle increased thermal density
Microfluidic Cooling is one Solution



We still must handle increased thermal density Microfluidic Cooling is one Solution


T. Brunschwiler et al., 3D-IC 2009 (IBM)

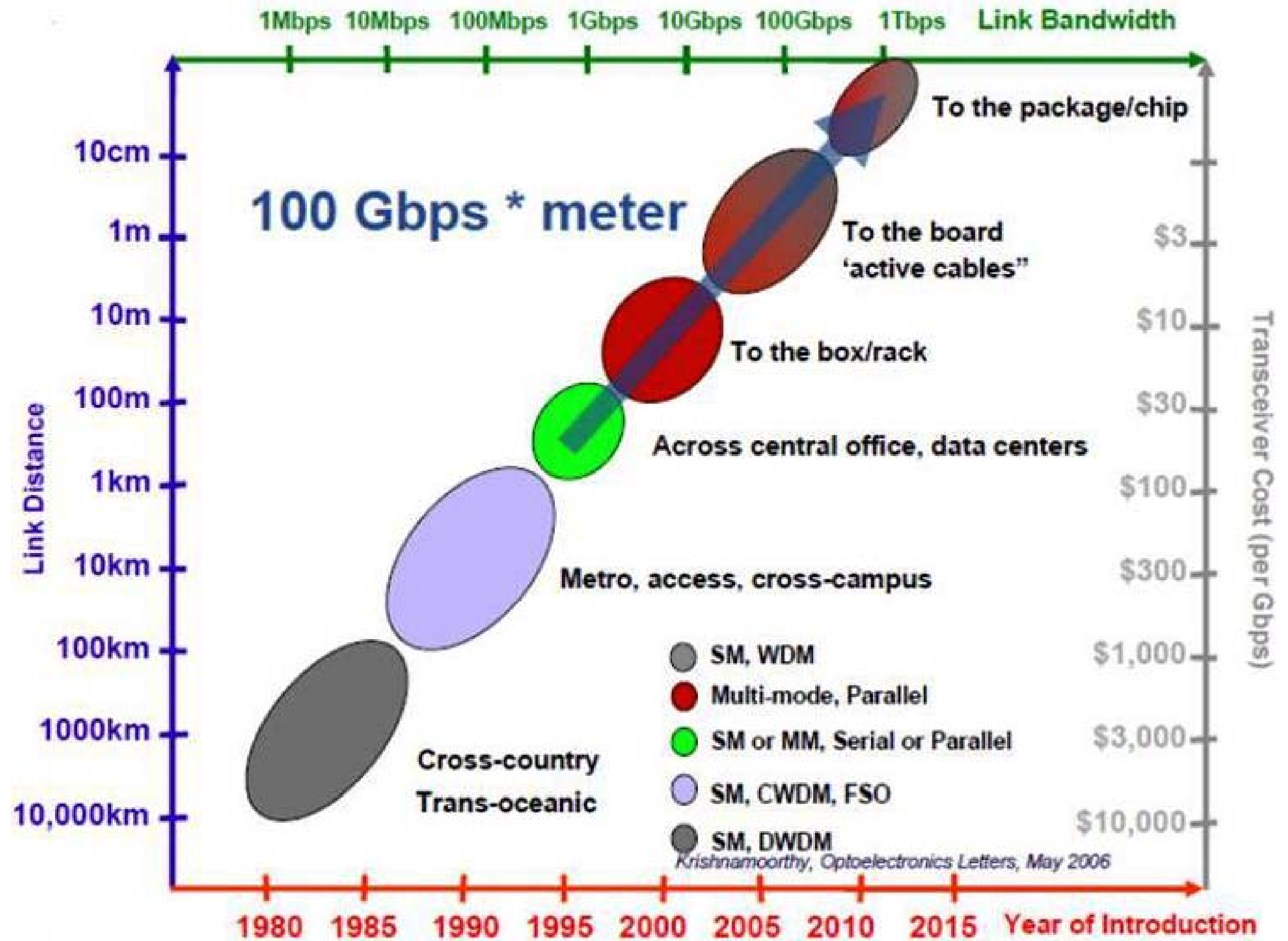
Reduce semiconductor power

- Reduce leakage currents (1.2X)
- Reduce on-chip Interconnect power by:
  - Improved conductor conductivity (TBD)
  - Decrease capacitance (1.33X)
- Reduce interconnect length (10 layers = 3.2X)
- Reduce operating frequency (TBD 1.2X)
- Reduce operating voltage (1.6 to 0.4V=4X V^2= 16X)
- Voltage regulator/core (use case dependent 1.1X)
- Reduce high speed electrical signal length (TBD)

(108X)

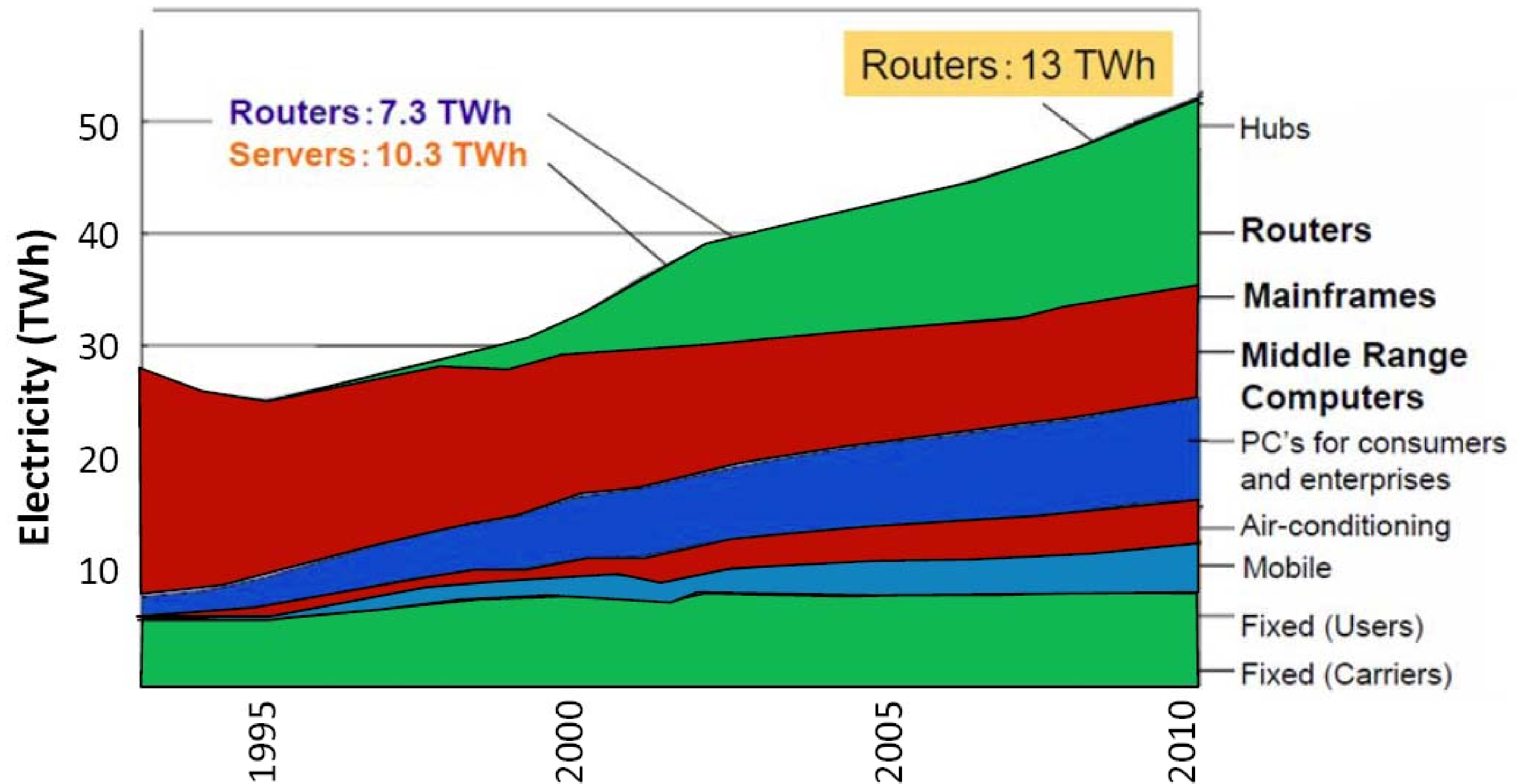
The “low handing” Fruit

- **Move the photons as close to the transistors as possible**



The “low handing” Fruit

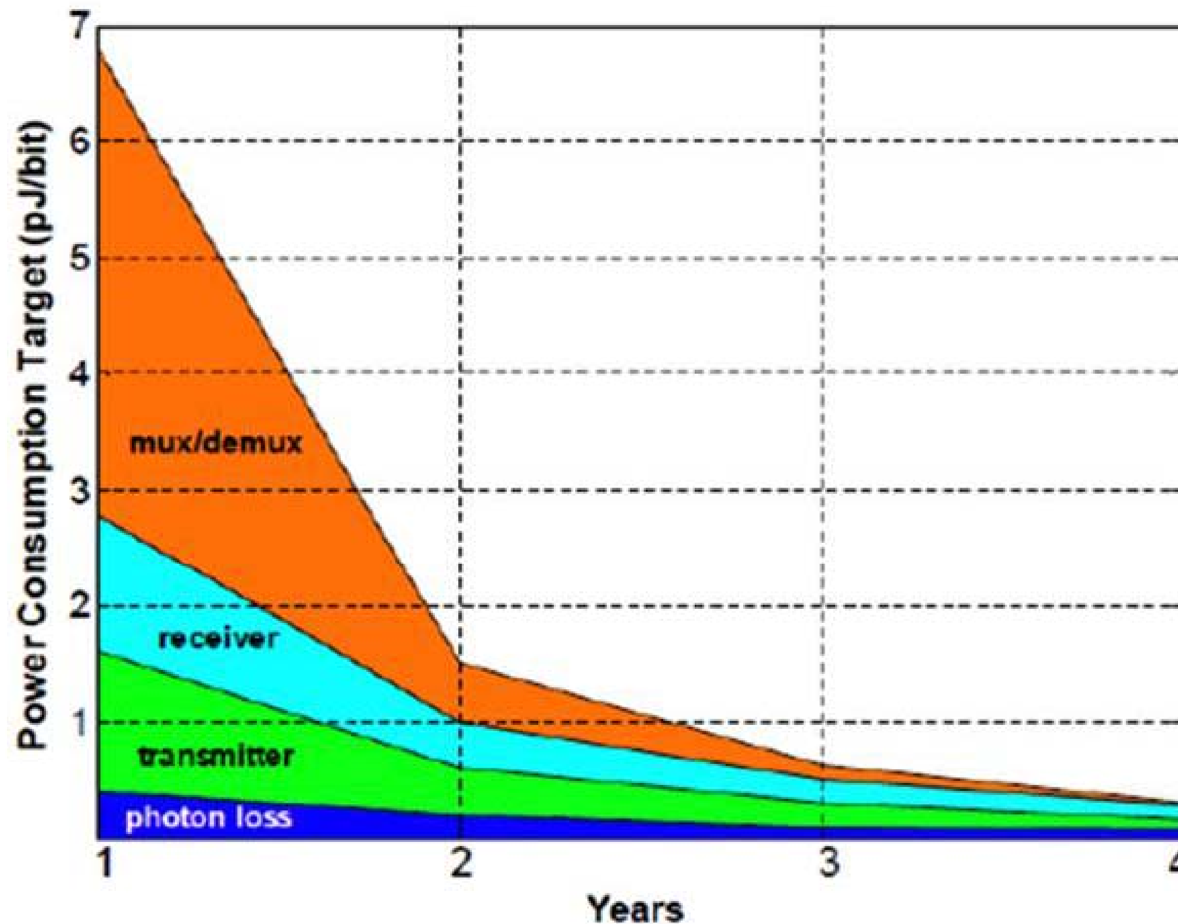
- **Reduce power in servers and routers**



Based on T. Asami and S. Namiki, ECOC 2008, Belgium, 2008

The “low handing” Fruit

- **Reduce power in the serdes circuits**

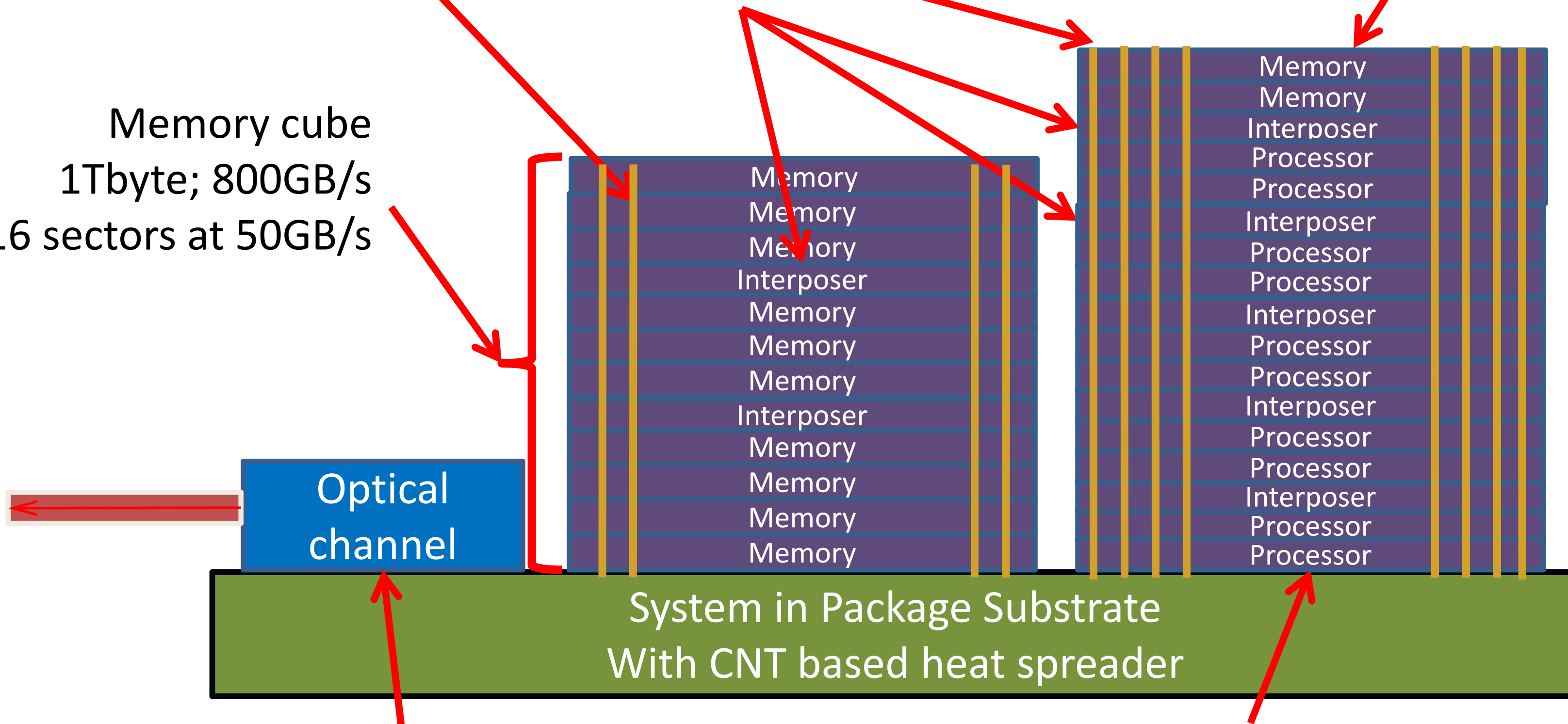


The 100X solution by 2015

TSV die to die connection

Silicon Interposer with:
-Integrated thermal management
-Integrated Passive networks

200GB; 800GB/s memory
(16 sectors at 50GB/s)



2TB/s optical transceiver for:
- Off package communication
- On package routing

Processor with 1000 cores/10 layers
Core transistor speed 1GHz
25um thick wafer (~400mV power)

Thank You