SAND 2008-7522C

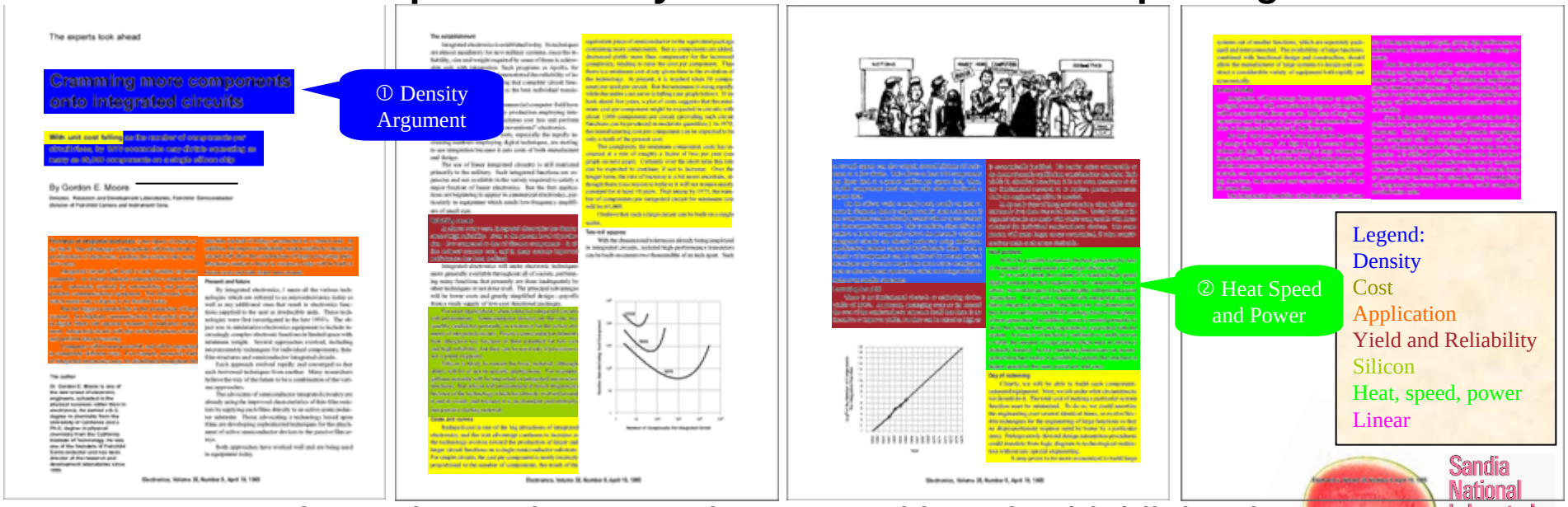
HPC Performance Limits and CMOS Maturity

Erik P. DeBenedictis
Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the
United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

The New Moore's Law

- As originally defined, Moore's Law related density, speed, and power such that μP throughput doubled every 18-20 months
  - The computer industry's infrastructure is based on this
- Speed and power "dropped off the curve" 2003-2008 and a new Moore's Law has density scaling only
  - The computer industry's infrastructure needs updating



Moore's Paper from Electronics magazine 1965 with topics highlighted

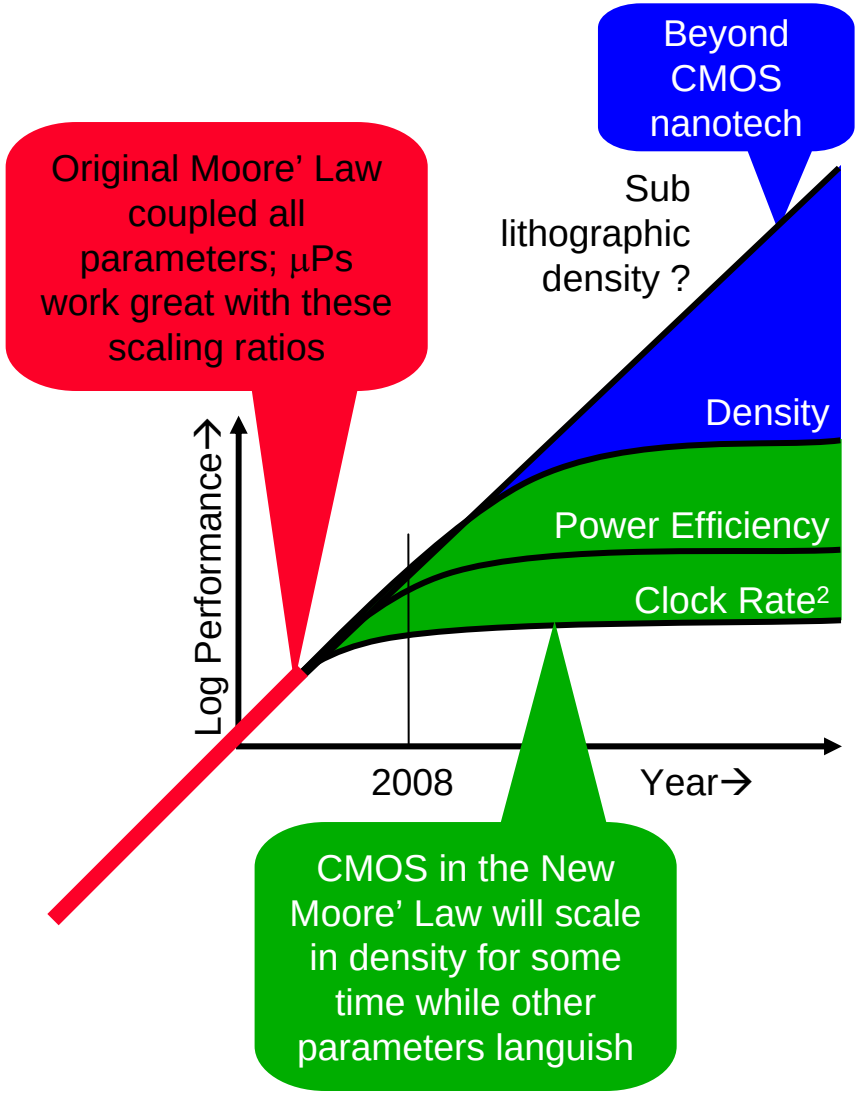
Sandia National Laboratories

CMOS Scaling in the Era of Moore's Paper and Now

- Dennard Scaling
  - Area $1/\kappa^2$
  - Capacitance $1/\kappa$
  - Resistance κ
  - Threshold (V_{th}) $1/\kappa$
  - Current (I_d) $1/\kappa$
  - Gate Delay (τ_{gd}) $1/\kappa$
  - Wire Delay (τ_{wire}) 1
  - Power $1/\kappa^2 \rightarrow 1/\kappa^3$

- Moore understood basic scaling rules
- Rules to left are Dennard scaling rules based on parameter κ
- Dennard scaling held for decades, which fostered the μP and a lot of computing infrastructure
- No longer holds due to multiple effects
  - Lack of wire delay scaling caught up
  - Microprocessors can't be made more complex to get faster single thread throughput anymore

Moore's Law Environment



- **Scaling**
  - **Until about 2005, density, speed, and power scaled in a specific ratio**
    - **density ∝ speed² ∝ 1/power ∝ e^t**
  - **Clock rate nearly flat lined a few years ago and power-efficiency growth is falling behind expectations**
  - **Lithographic density will flat line in a decade per ITRS**
  - **Sub lithographic density may continue density scaling for a long time (but don't really know)**
- **Objectives**
  - **Understand CMOS scaling to end game**
    - **Three curves on graph**
  - **Understand CMOS + sub lithographic nanotechnology**
    - **All four curves on graph**

Sandia National Laboratories

Key Problems

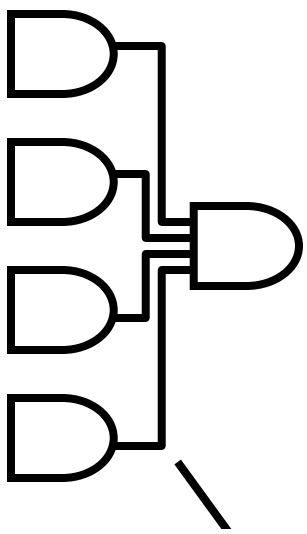
- **Speed. Clock rate limits can be mitigated with parallelism, but hardly anybody knows how to program in parallel.**
  - **Except this group which invited me to talk comprises the world's experts on parallelism and is asking me to skip this issue.**
- **Power. There are some serious power obstacles, but there is wiggle room.**

Power Limits

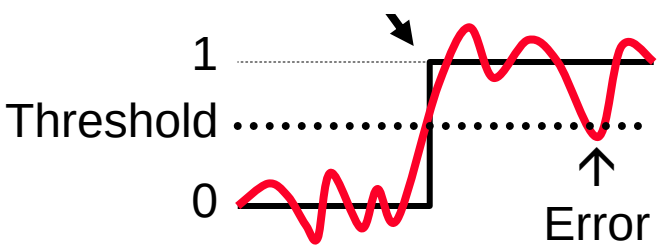
- Landauer's Limit $k_B T \ln(2)$
  - Device independent and applies to AND-OR-NOT logic
  - We cannot approach Landauer's Limit with CMOS, other limits arrive first
- Thermal Limit 100 $k_B T$
  - This might stop the current show (show stopper)
  - The glitch rate for a gate at N $k_B T$ energy is about $e^{-N} \approx 4\times10^{-44}$ for N=100. For N < 100 a computer would start to become unreliable. The limit is soft and depends on user expectations of reliability.

Thermal Limit Details

Test Circuit



Waveform with Thermal Noise



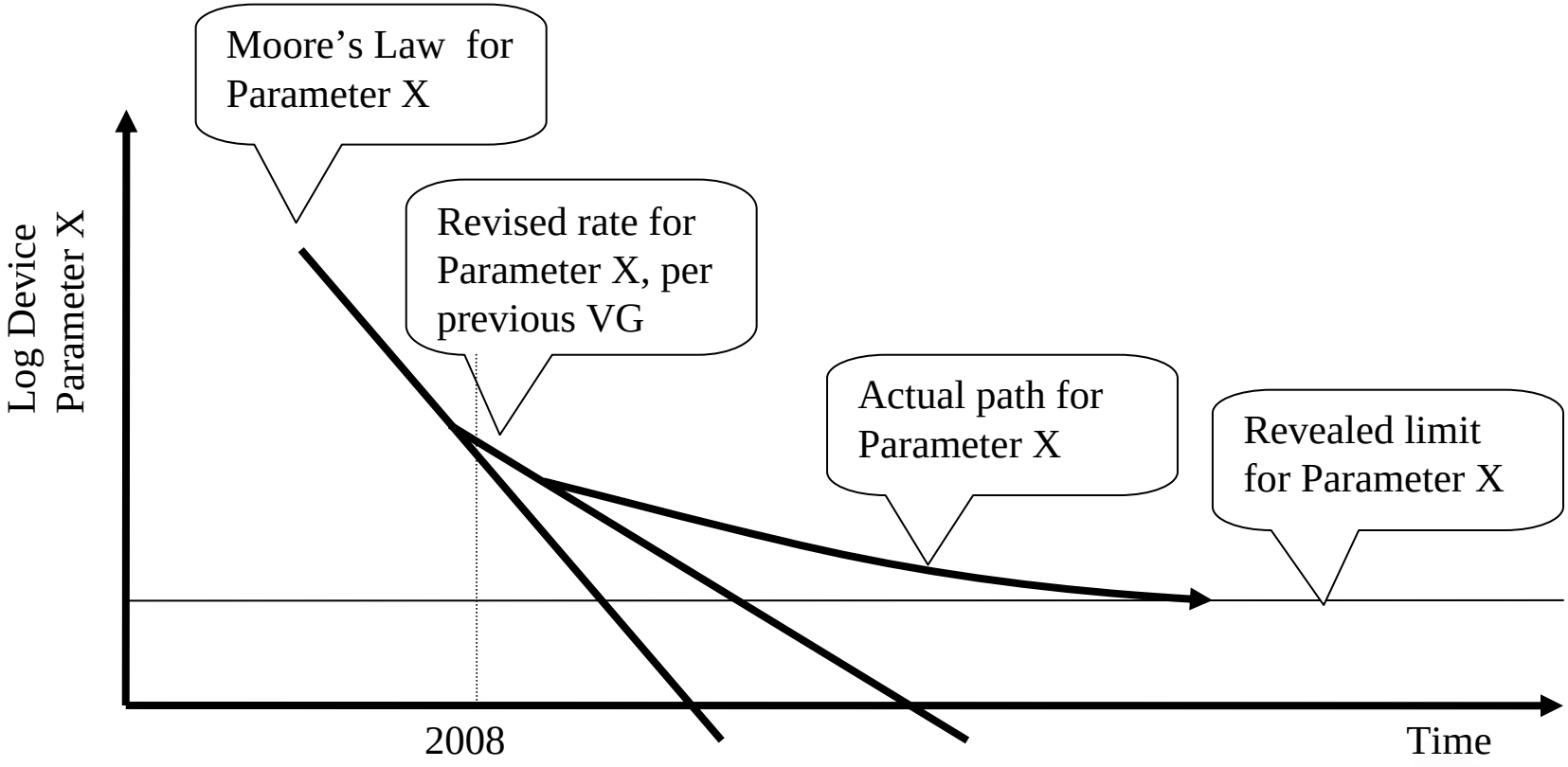
- **Described electrically, but applies generally**
- **Details**
  - **Noise power = $k_B T \omega$**
    - **k_B Boltzmann's constant**
    - **T temperature**
    - **ω bandwidth**
  - **An efficient design clocks every $1/\omega$ seconds**
  - **Noise power per clock period is $k_B T$**
  - **If signal energy is $Nk_B T$, probability that noise exceeds clock is about e^{-N}**

Sandia National Laboratories

Semiconductor Roadmap

- CMOS Limits
  - Some CMOS advocates propose scaling to the 100 k_BT limit and below
    - In fact, some go below Landauer's Limit
  - However, graphs are shifting slope suggesting lower limits
  - Analogy: Steam power plants top out at around 40% efficiency

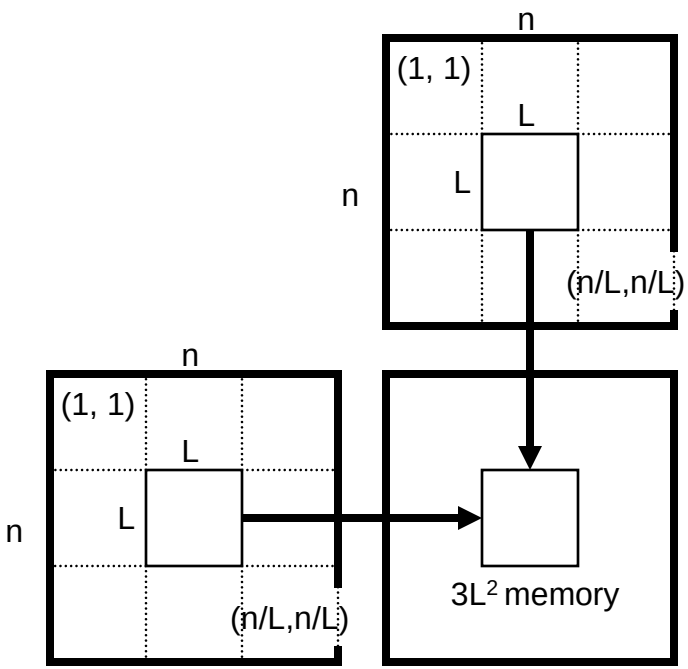
Interpretation of Graph



Mitigations

- Algorithms
- Software → Hardware
- Architecture
- Upside Potential

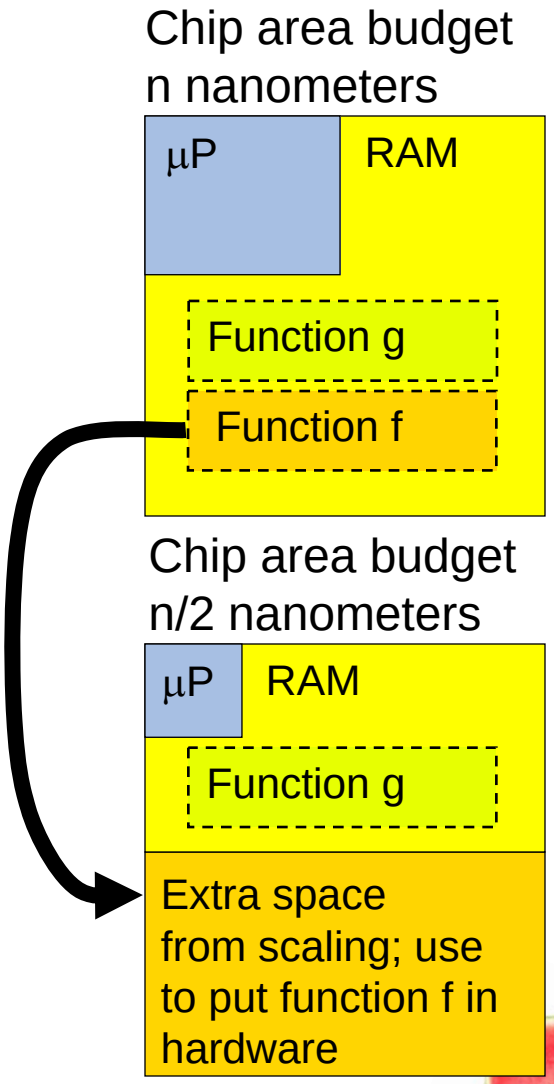
Using Memory to Save Power



- Consider processor that can hold $3L^2$ numbers
- Algorithm loads $L \times L$ blocks onto the processor, performs the L^3 complexity local matrix multiply, then stores one $L \times L$ block
- Memory accesses are $(n/L)^3\ L \times L = n^3/L$
  - Memory $\propto n^3/L$
- Key point: Memory subsystem power can be reduced by effective use of memory
  - Memory will continue to scale up, but power is flat lining
  - Proposal is to study algorithms of this type generally

Translating Software to Hardware

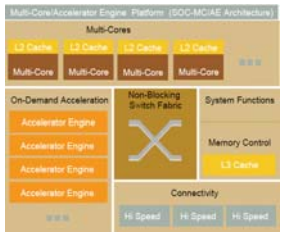
- **Software is space efficient but power inefficient, which is exactly opposite of the direction for progress**
- **Physical device scaling will give us an enormous rise in device count, but we will overheat the chip if we use the extra devices for more microprocessors (cores)**
- **Remedy is to shift functions from space-efficient to power-efficient implementations**
  - **Hardware implementations can be highly parallel (fast) and power-efficient**
  - **However, they may need to run at a low duty cycle**
- **Example shows 2:1 scaling; project objective is architectures that scale this way over a continuous range**



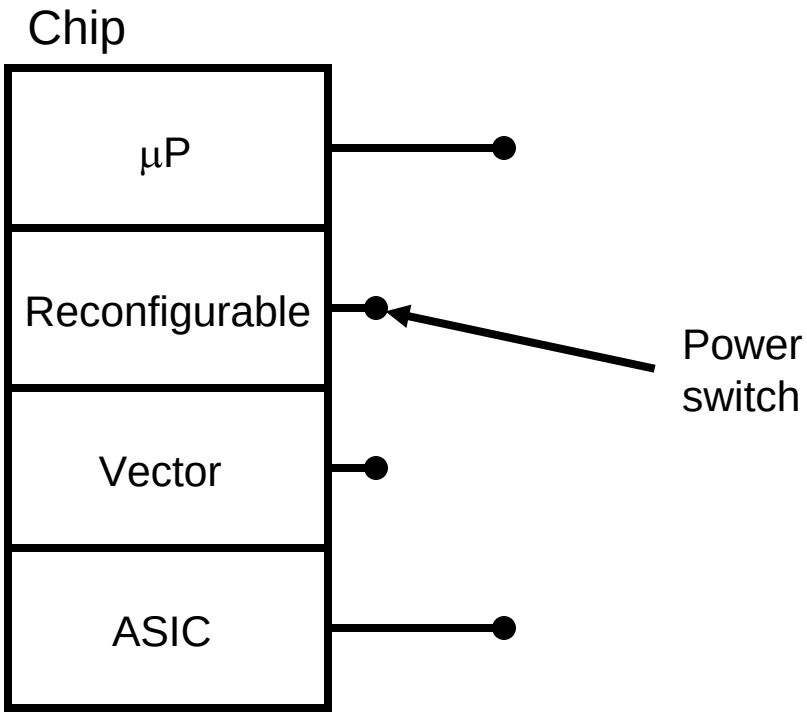
Sandia National Laboratories

Hybrid Multi-Architecture

- **Background**
  - **A μP is general but power inefficient**
  - **Other architectures are less general but more power efficient**
- **Opportunity**
  - **Chips have multiple architectures**
  - **Functions are moved to whichever architecture executes them with best power efficiency**
  - **Raises power efficiency through use of more device count. This is the right path to the future**
- **Research Program**
  - **Find architectures of this class for DoD programs**
  - **Find architectures of this class that also include nanotech architectures from other parts of the program**
- **Note: Industry does this too →**



2007 ITRS System Drivers p. 4



Rotate switch to power up only the architecture needed at a given instant

Sandia National Laboratories

Architecture

- Emerging processors deploy multiple special purpose units, of which not all are powered up at any given time
- Examples:
  - CPU+GPU in desktop
  - Roadrunner CPU + Cell
  - Future proposals of industry such as CPU + GPU + encryption + cell phone radio

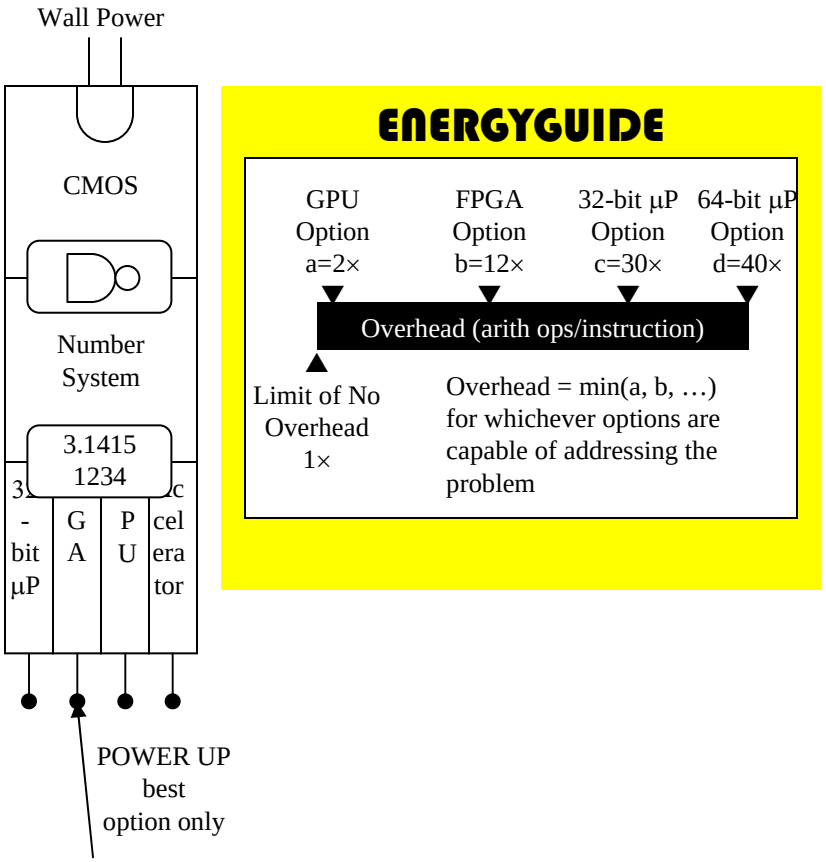


Figure : Hybrid Architecture. This is a generalization of many activities in the industry (see text). The processor stage comprises several options with different efficiency for different tasks. Through some form of power management, one unit carries the computational burden for a task while the others are powered down or left in a static standby state.

Conclusions

- Semiconductor scaling changed ~2005
  - "original Moore's Law" gives way to
  - "new Moore's Law," lacking speed & power scaling
- The μP no longer natural fit (eigenvector) of semiconductor scaling
- Top problem for general community: speed
- Top problem for parallel programmers: power
- Remedies
  - Special purpose and multiple architectures
  - Power-efficiency improving algorithms
  - Slow down the clock