CSE 548: Computer Systems Architecture

Technology

Spring 2017

Luis Ceze (Instructor)
Thierry Moreau (TA)

(slides from many friends, most specially Milo Martin (UPenn) and Krste Asanovic (UC Berkeley).

Architecture triangle

"Technology"
Logic Gates
SRAM
DRAM
Circuit Techniques
Packaging
Magnetic Storage
Flash Memory

Constraints
Function
Performance
Reliability
Cost/Manufacturability
Energy Efficiency
Time to Market
Form factor

Domains
Desktop
Servers/cloud
Tablets/Mobile Phones
Supercomputers
Game Consoles
Embedded/IoT

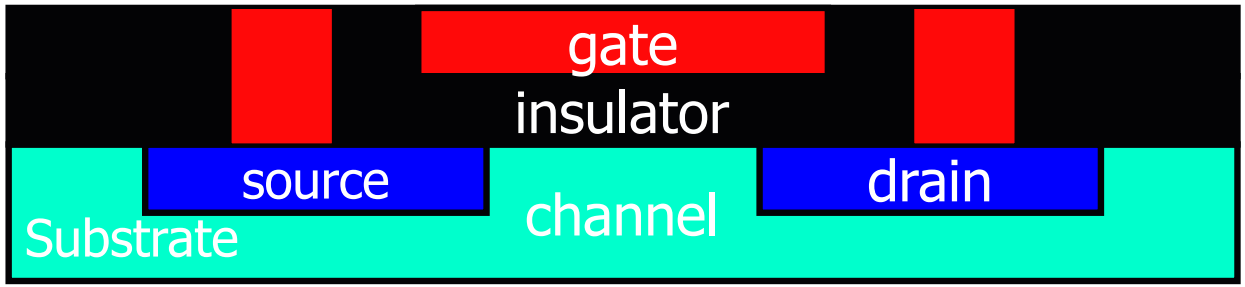
Aside: Better computers help design the next generation (CAD)!

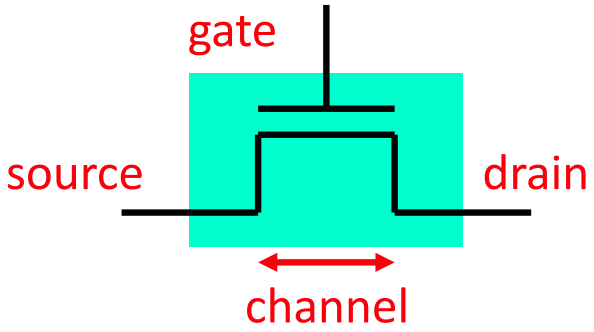
The dominant landscape: Tablets/phones/IoT backed by Warehouse-scale computers

What do I mean by "Technology"

- Basic element
  - Solid-state **transistor** (i.e., electrical switch)
  - Building block of **integrated circuits (ICs)**
- What's so great about ICs? Everything
  - \+ High performance, high reliability, low cost, low power
  - \+ Lever of mass production
- Several kinds of IC families
  - **SRAM/logic**: optimized for speed, used for processors
  - **DRAM**: optimized for density, cost, power, used for memory
  - **Flash**: non-volatile memory
  - Increasing opportunities for integrating multiple technologies
- Non-transistor storage and inter-connection technologies
  - Disk, optical storage, ethernet, fiber,

Semiconductor Transistor

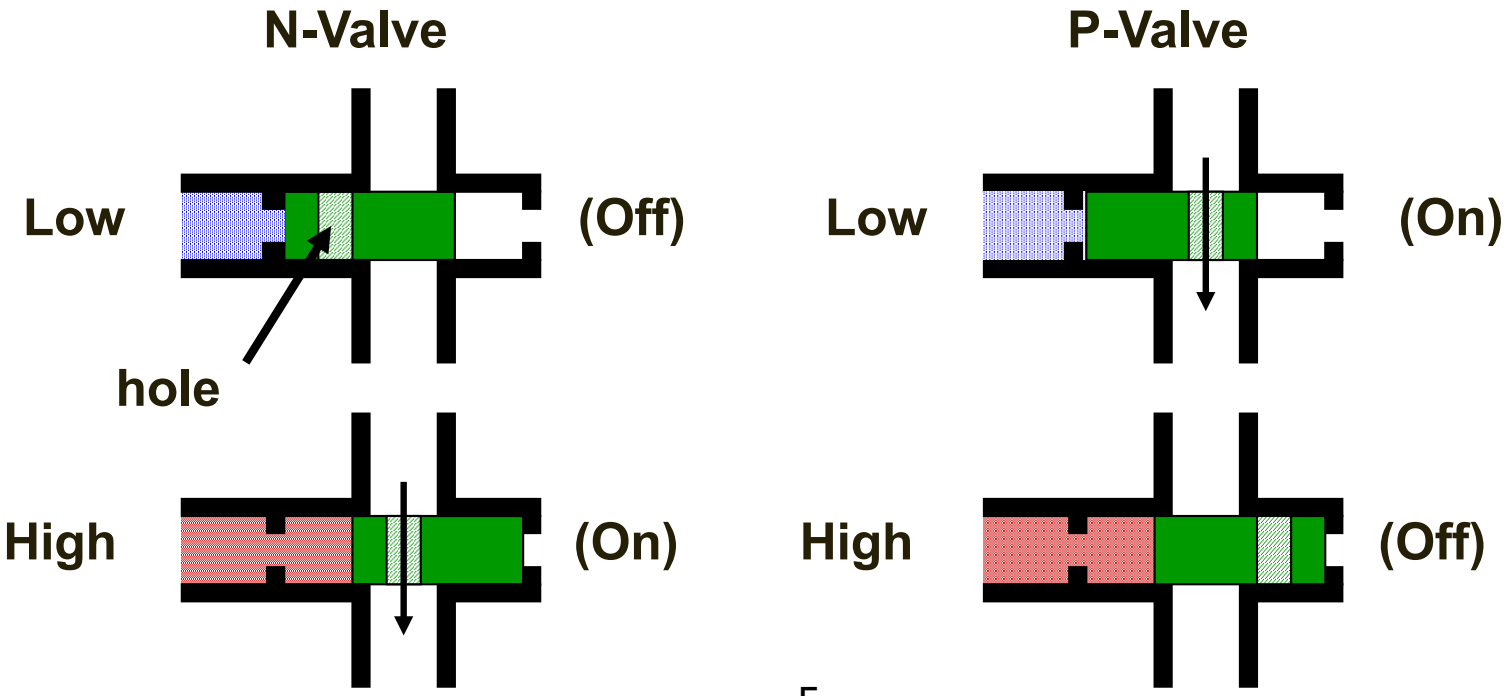


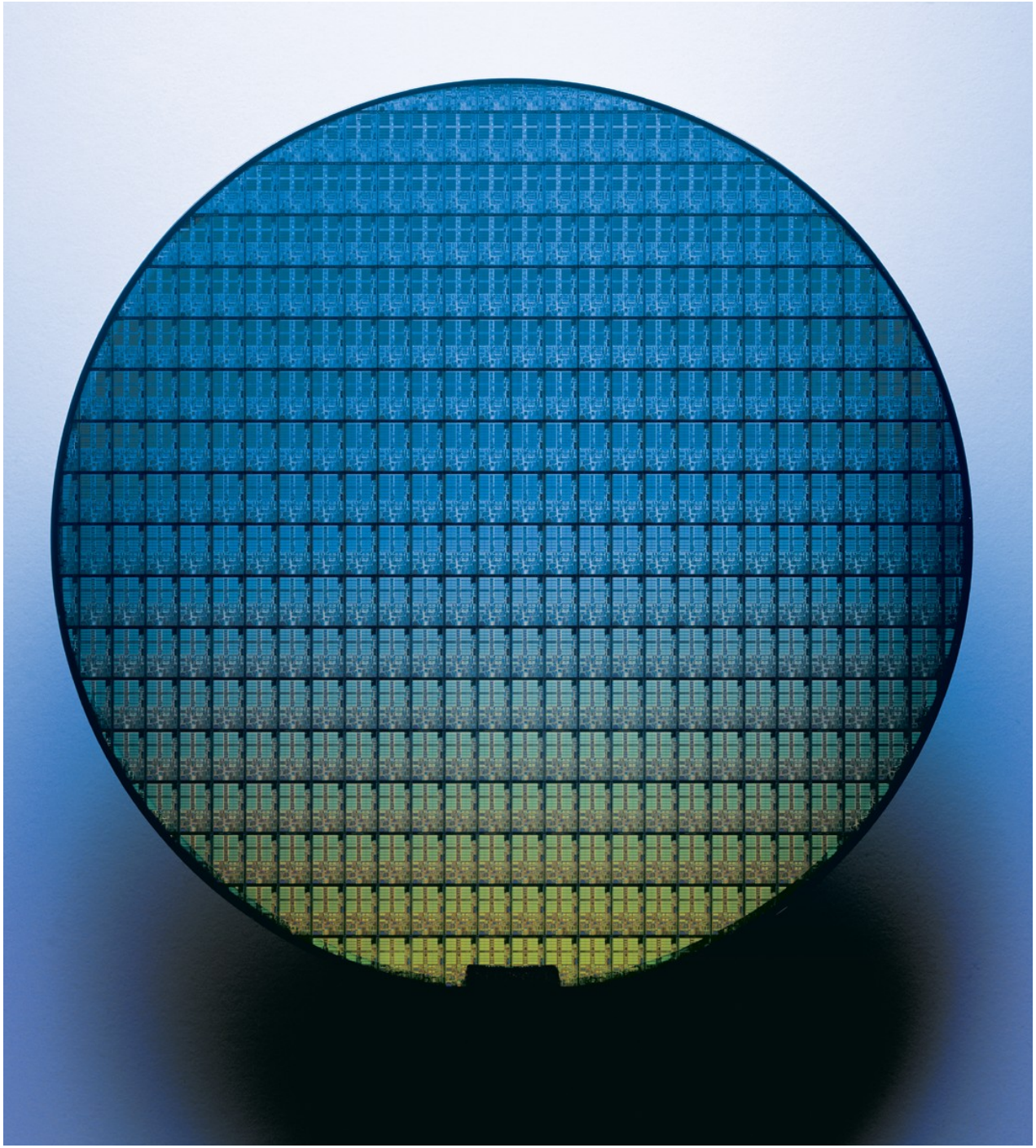


- Basic technology element: **MOSFET**
  - Solid-state component acts like electrical switch
  - **MOS**: metal-oxide-semiconductor
    - Conductor, insulator, semi-conductor
- **FET**: field-effect transistor
  - Channel conducts source→drain only when voltage applied to gate
- **Channel length**: characteristic parameter (short → fast)
  - Aka "feature size" or "technology node"
  - Currently: 14 nanometers (nm)
  - Continued miniaturization (scaling) known as "**Moore's Law**"
    - Won't last forever, physical limits approaching (or are they?)

A Transistor Analogy: Computing with Air

- Use air pressure to encode values
  - High pressure represents a "1" (blow)
  - Low pressure represents a "0" (suck)
- Valve can allow or disallow the flow of air
  - Two types of valves

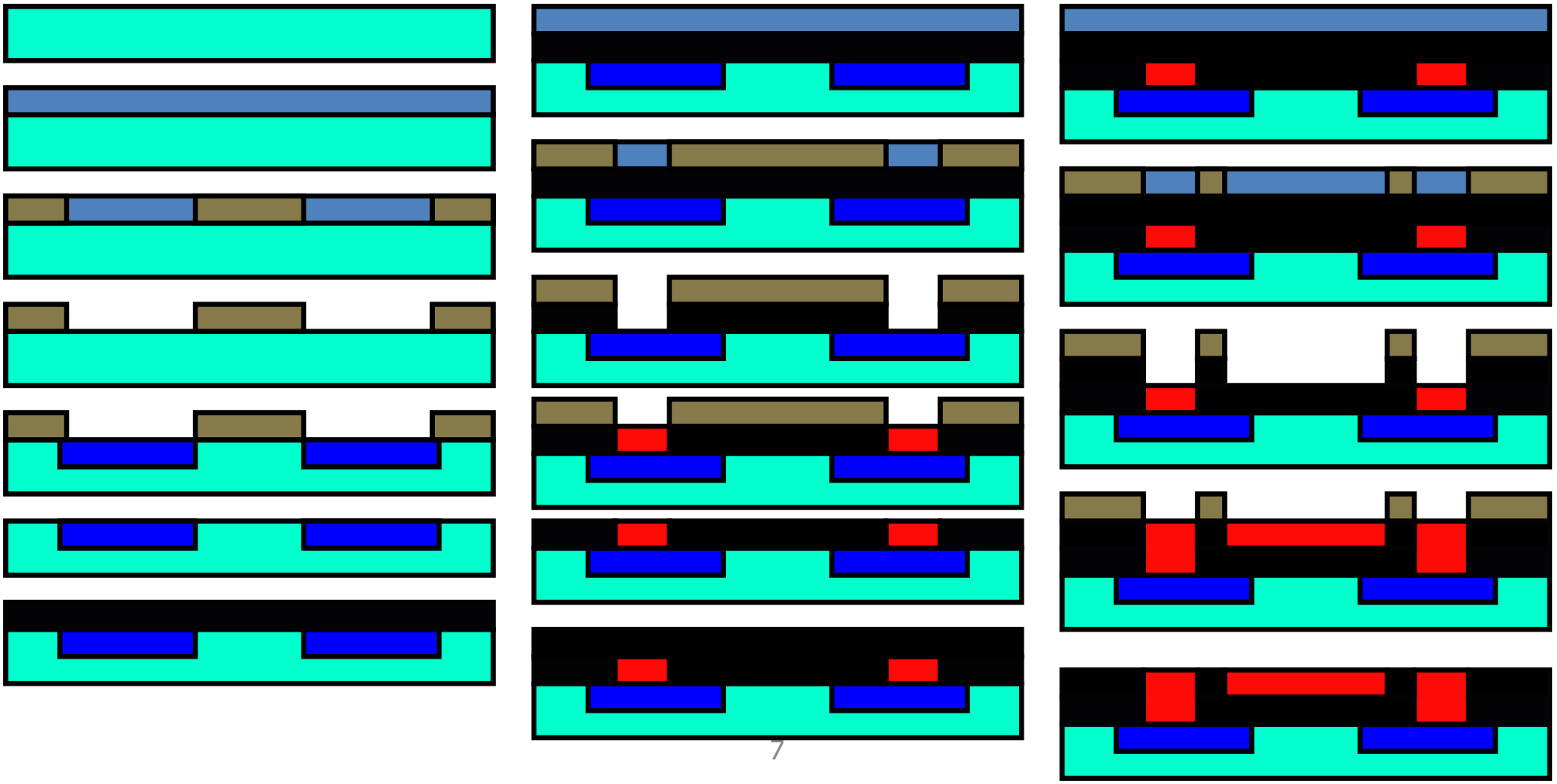




Intel
Pentium M
Wafer

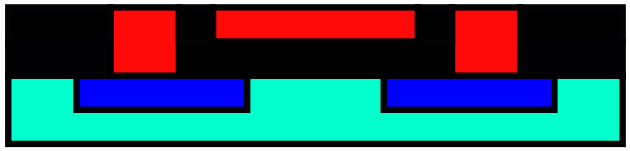
Manufacturing Steps

- Multi-step photo-/electro-chemical process
  - More steps, higher unit cost
- \+ Fixed cost mass production ($1M+ for “mask set”)

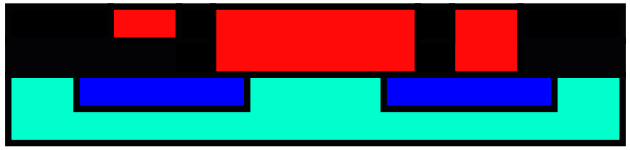


Manufacturing Defects

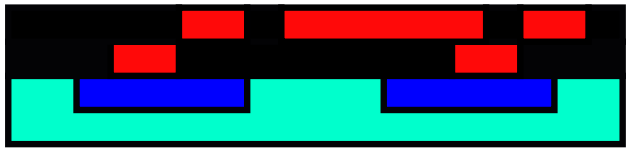
Correct:



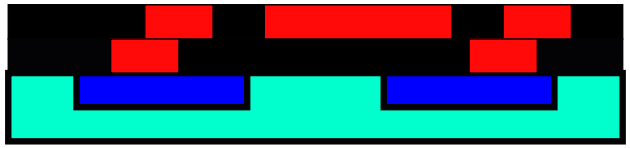
Defective:



Defective:



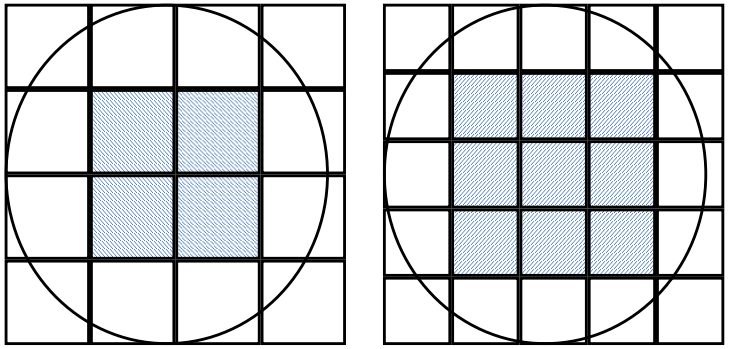
Slow:

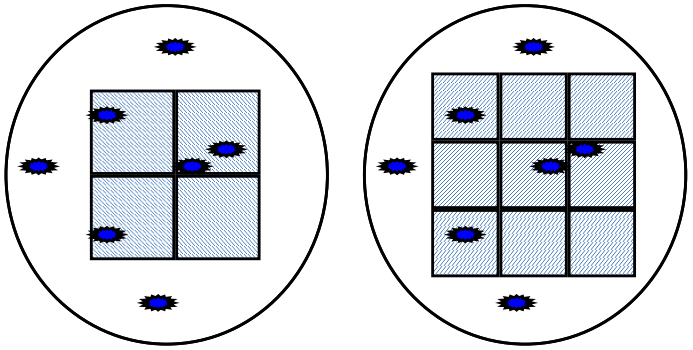


- Defects can arise
  - Under-/over-doping
  - Over-/under-dissolved insulator
  - Mask mis-alignment
  - Particle contaminants
- Try to minimize defects
  - Process margins
  - Design rules
    - Minimal transistor size, separation
- Or, tolerate defects
  - Redundant or “spare” memory cells
  - Can substantially improve yield

Cost Implications of Defects

- Chips built in multi-step chemical processes on **wafers**
  - Cost / wafer is constant, f(wafer size, number of steps)
- Chip (die) cost is related to **area**
  - Larger chips means fewer of them
- Cost is **superlinear** in area
  - Why? random defects
  - Larger chip, more chance of defect
    - Result: lower "yield" (fewer working chips)

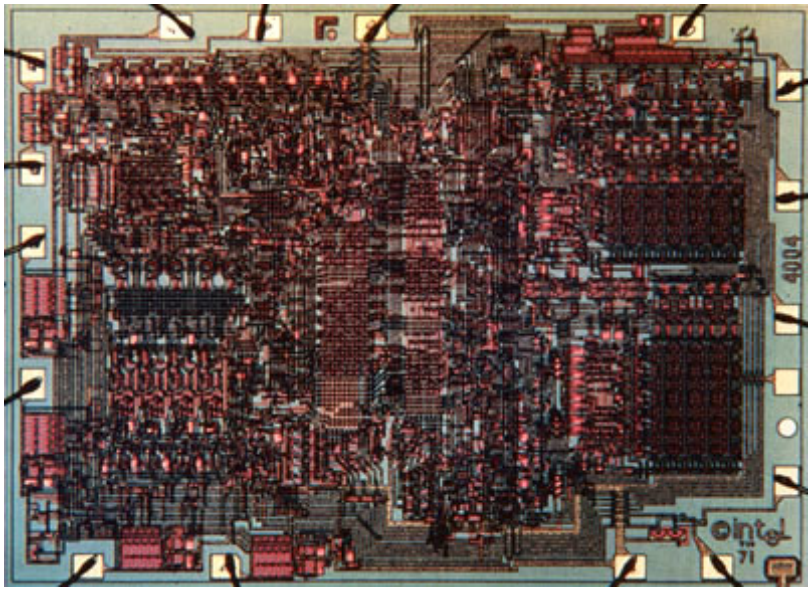




- **Wafer yield**: % wafer that is chips
- **Die yield**: % chips that work
- Yield is increasingly non-binary - fast vs slow chips

First Microprocessor

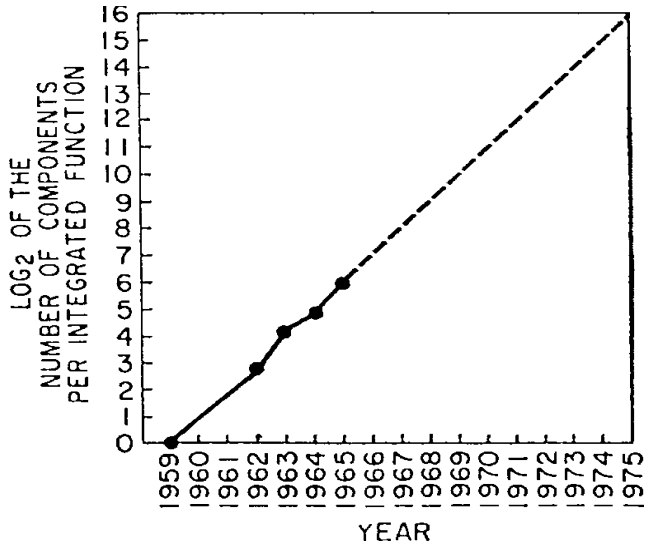
- Connect a few transistors together to make…



- Intel 4004
  - 1971 (first microprocessor)
  - 4-bit data
  - 2300 transistors
  - 10 μm technology
  - 108 KHz
  - 12 Volts
  - 13 mm^2
  - 20 KIPS (thousand instructions per second)

Moore's Law (1965)

- Transistors per inch square
  - Twice as many after ~1.5-2 years
- Some technology-based ramifications
  - Annual improvements in density, speed, power, costs
  - SRAM/logic: density: ~30%, speed: ~20%
  - DRAM: density: ~60%, speed: ~4%
  - Disk: density: ~60%, speed: ~10% (non-transistor)
  - Big improvements in flash memory and network bandwidth, too
- Related trends
  - Processor performance
    Twice as fast after ~18 months
  - Memory capacity
    Twice as much in <2 years
- **Changing quickly and with respect to each other!!**
  - Example: density increases faster than speed
  - Trade-offs are constantly changing
  - **Re-evaluate/re-design for each technology generation**
- Reading: Moore's original paper



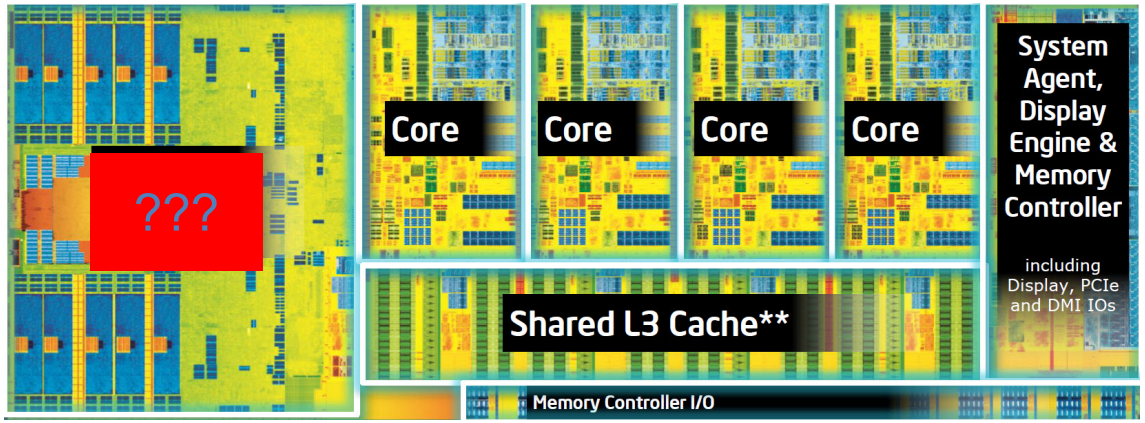
Today:
2^{32} transistors

How were growing # of transistors used?

- Initially to widen the datapath
  - 4004: 4 bits (BCD calculators) $\rightarrow$ Pentium4: 64 bits
- … and also to add more powerful instructions
  - To amortize overhead of fetch and decode
  - To simplify programming (which was done by hand then)
- And?…

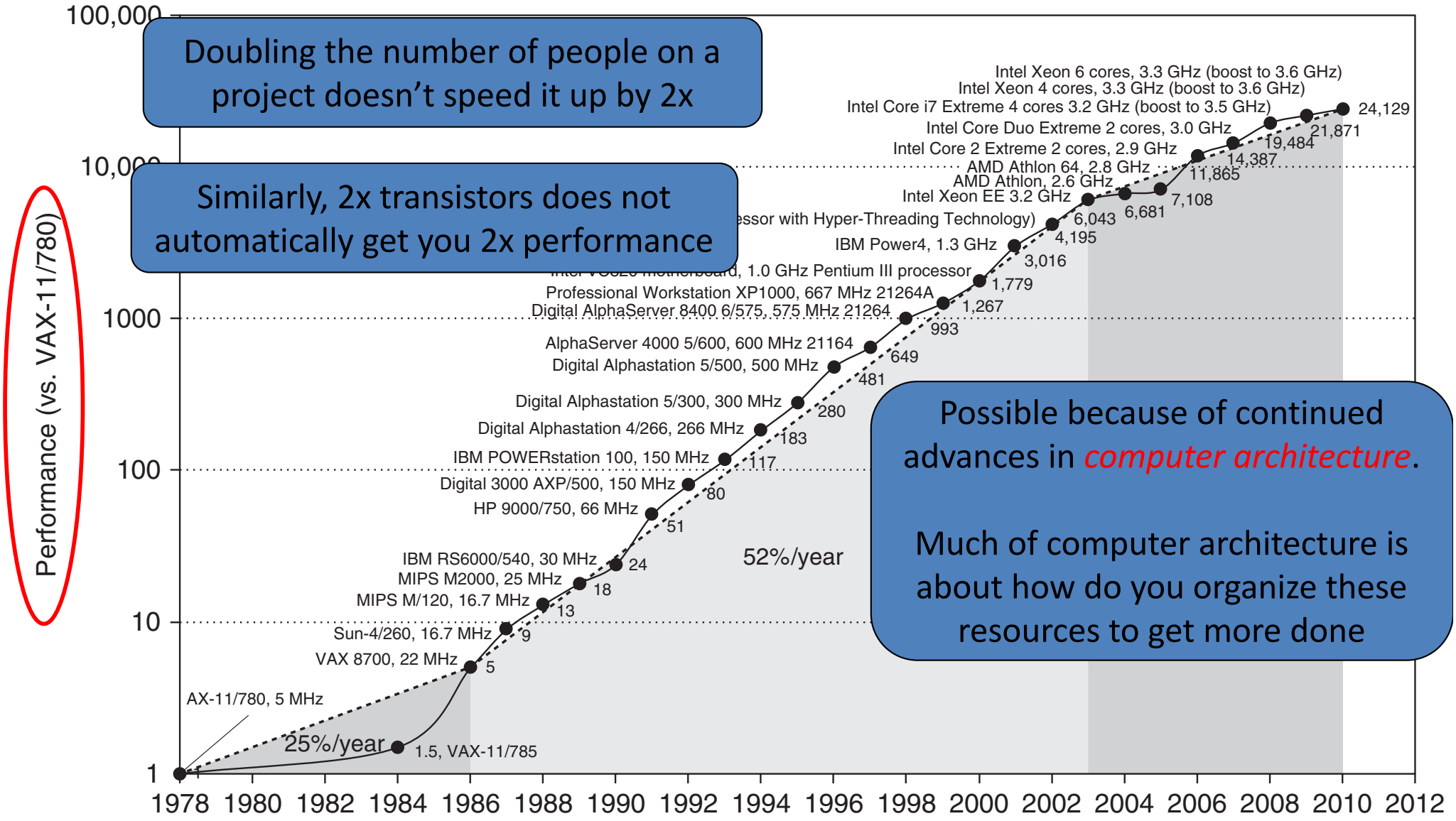
“Recent” Microprocessor

- Intel Core i7 (2013)
  - Application: desktop/server
  - Technology: 22nm (25% of P4)
  - 1.4B transistors (30x)
  - 177 mm^2 (2x)
  - 3.5 GHz to 3.9 Ghz (~1x)
  - 1.8 Volts (~1x)



  - 256-bit data (2x)
  - 14-stage pipelined datapath (0.5x)
  - 4 instructions per cycle (1x)
  - Three levels of on-chip cache
  - data-parallel vector (SIMD) instructions, hyperthreading
  - **Four-core multicore** (4x)

Performance Trend



The “Meter” of Computer Architecture

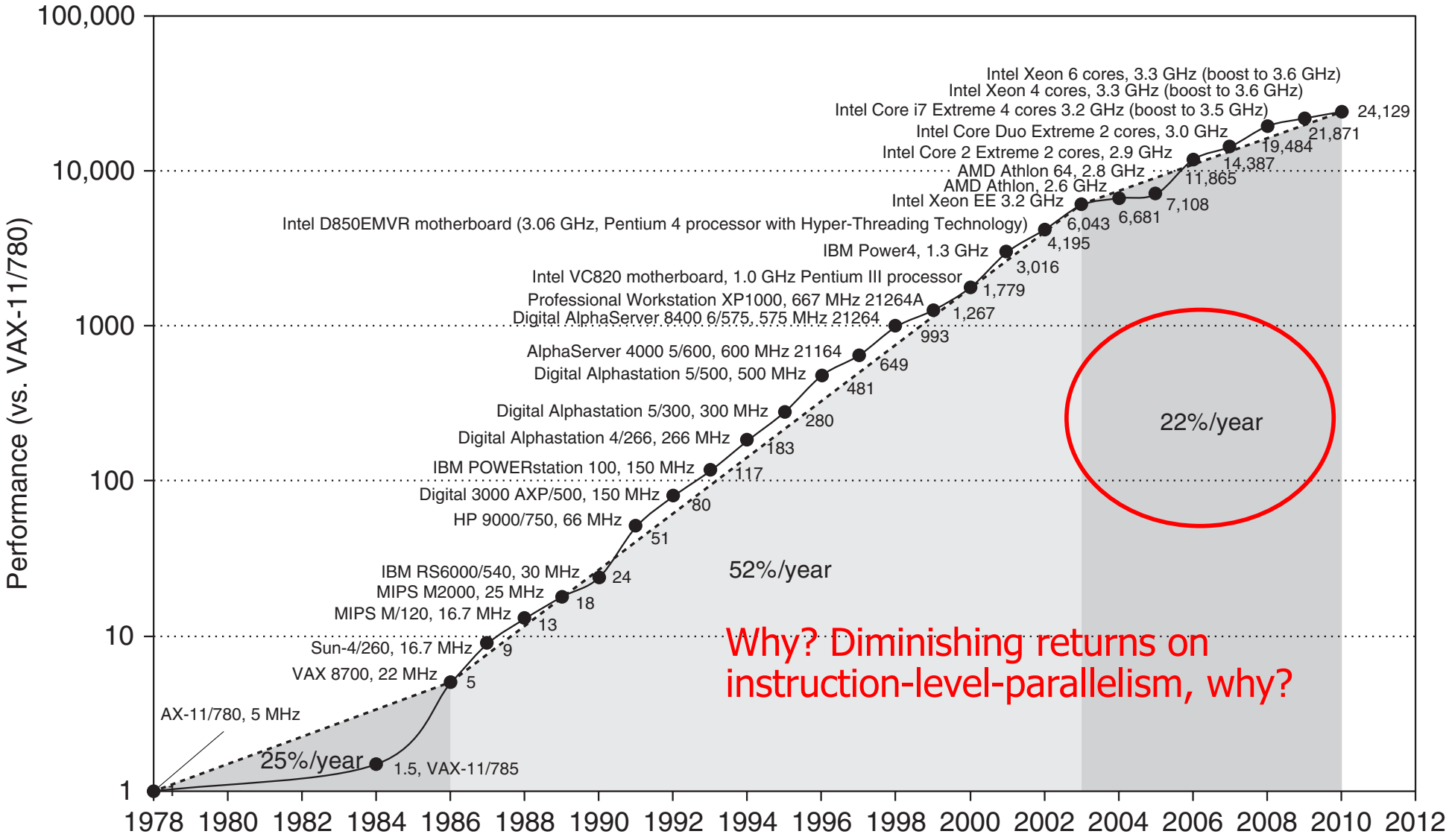
Implicit Parallelism

- **Extract implicit instruction-level parallelism**
  - Hardware provides parallel resources, figures out how to use them
  - Software is oblivious

- Initially using pipelining …
  - Which also enabled increased clock frequency
- … caches …
  - Which became necessary as processor clock frequency increased
- … and integrated floating-point
- Then deeper pipelines and branch speculation
- Then multiple issue (superscalar)
- Then dynamic scheduling (out-of-order execution)

- We will talk about these things

Hmm, have you noticed it?



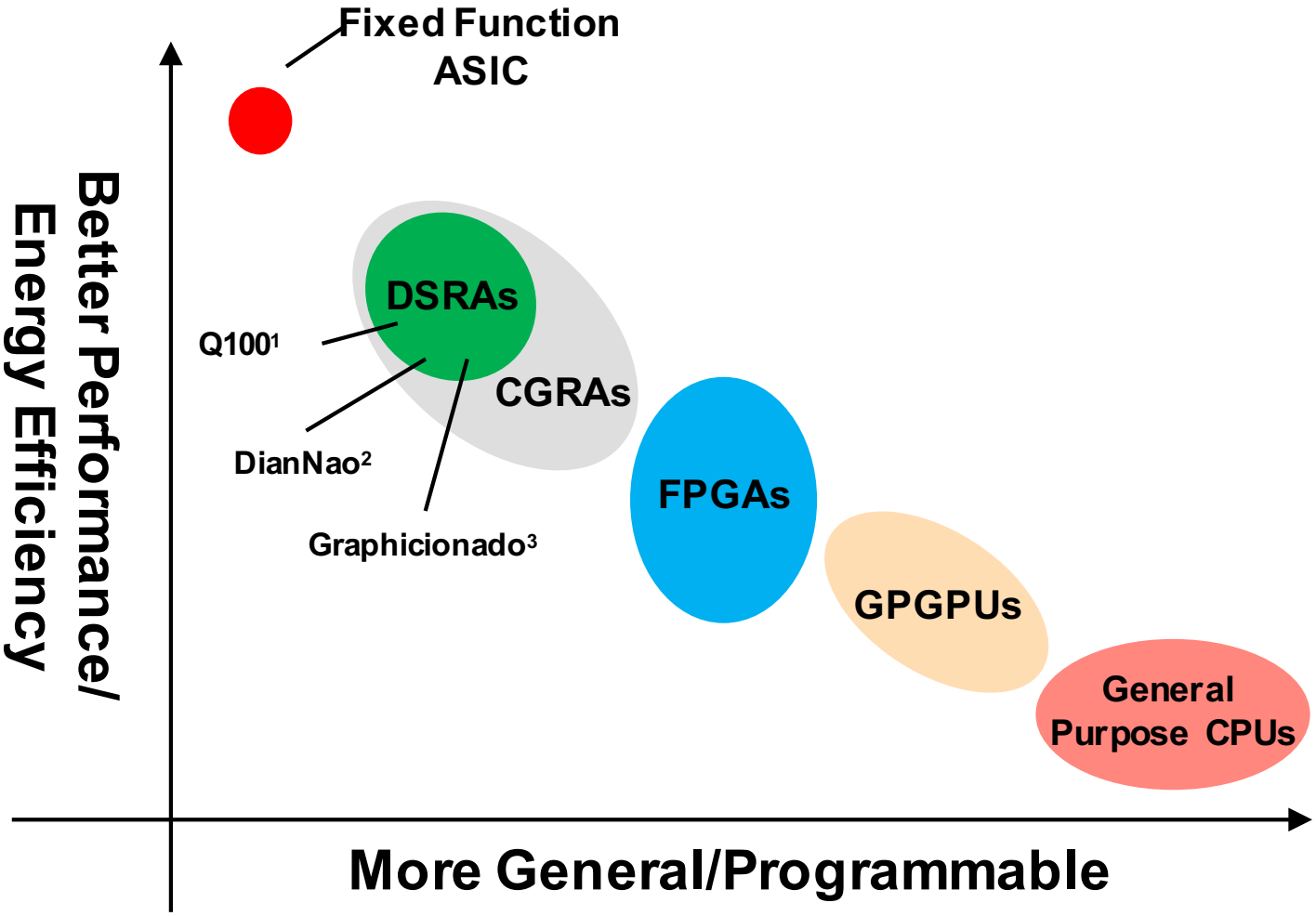
Explicit Parallelism

- Support **explicit data & thread level parallelism**
  - Hardware provides parallel resources, software specifies usage
- Helps alleviate power concerns – why?

- First using (subword) vector instructions…, Intel's SSE
  - One instruction does 4 parallel multiplies
- … and general support for multi-threaded programs
  - Coherent caches, hardware synchronization primitives
- Then using support for multiple concurrent threads on chip
  - First with single-core multi-threading, now with multi-core
- Integrated graphics? Accelerators? FPGAs?
  - AMD bought ATI, Nvidia making ARM procs, hmmm…

- (We will cover these too)

- Still not enough for sustained performance/energy efficiency improvement


Fixed Function
ASIC
DSRAs
Q100[1]
CGRAs
DianNao[2]
Graphicionado[3]
FPGAs
GPGPUs
General
Purpose CPUs
Better Performance/
Energy Efficiency
More General/Programmable