small processors solve BIG PROBLEMS

Chris Rowen
Founder and CTO
Tensilica, Inc.

10 May 2011

Agenda

- Four defining trends for the silicon universe
- Rethinking microprocessor design
  - World's shortest history of microprocessor design
  - Building microprocessors to order – configurability and extensibility
- Going ultra-small
- Some extreme processor applications
  - o A Server with Thousands of Processors per Chip?
  - o The World's Fastest DSP
  - o 1,000,000,000,000 RISC ops/s in 1mm^2 (Turbo Decoder)
  - o Putting Up To 250M Processors to Work - Exascale Climate Modeling
- Long-term Trends

Our New World

Mobile Broadband Terminals
(and sensors)

Network Infrastructure

The Computing Cloud

What's impressive:

- # users
- compute per watt

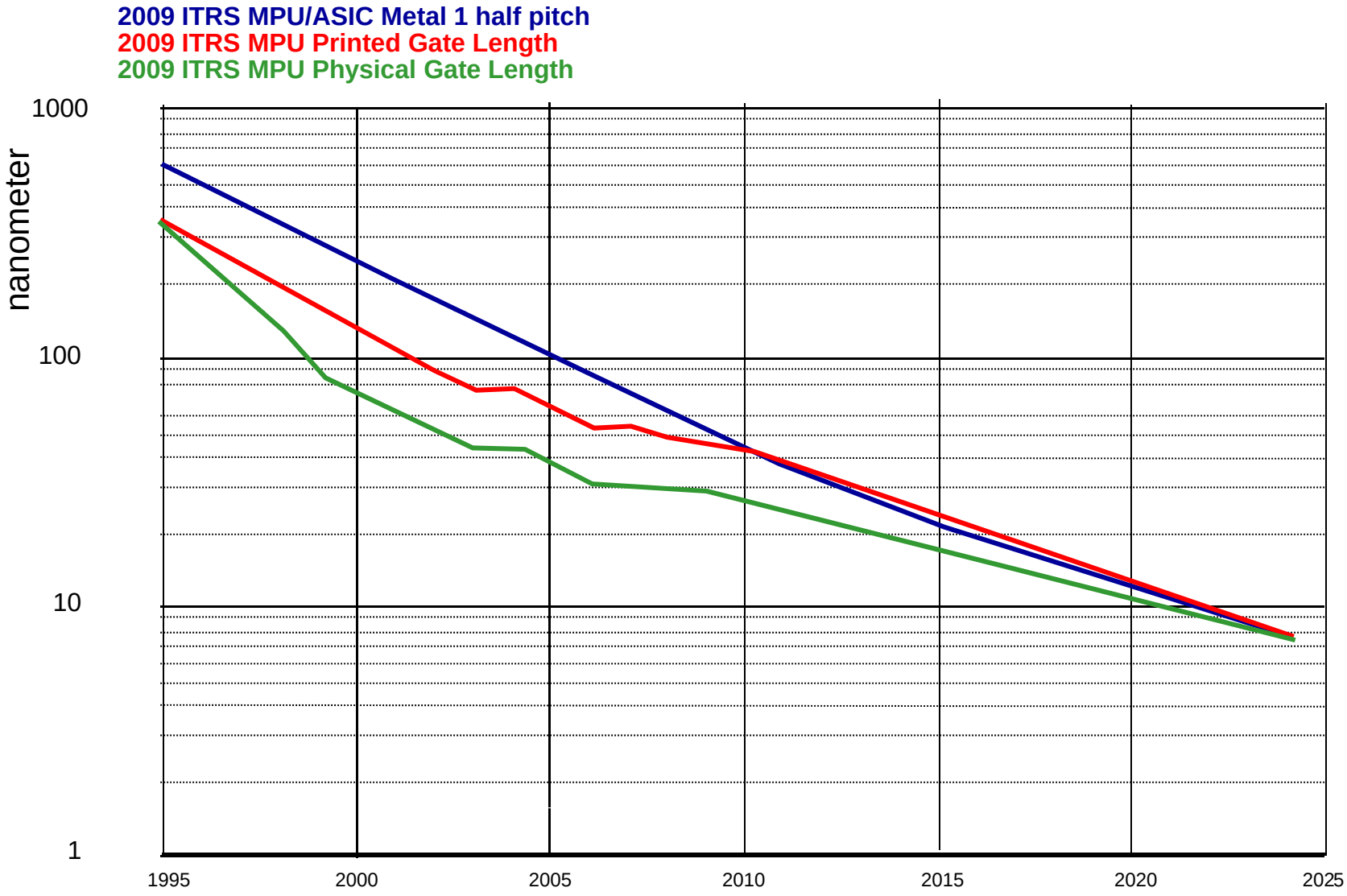
What's impressive:

- Aggregate bandwidth
- bits per Hz of spectrum

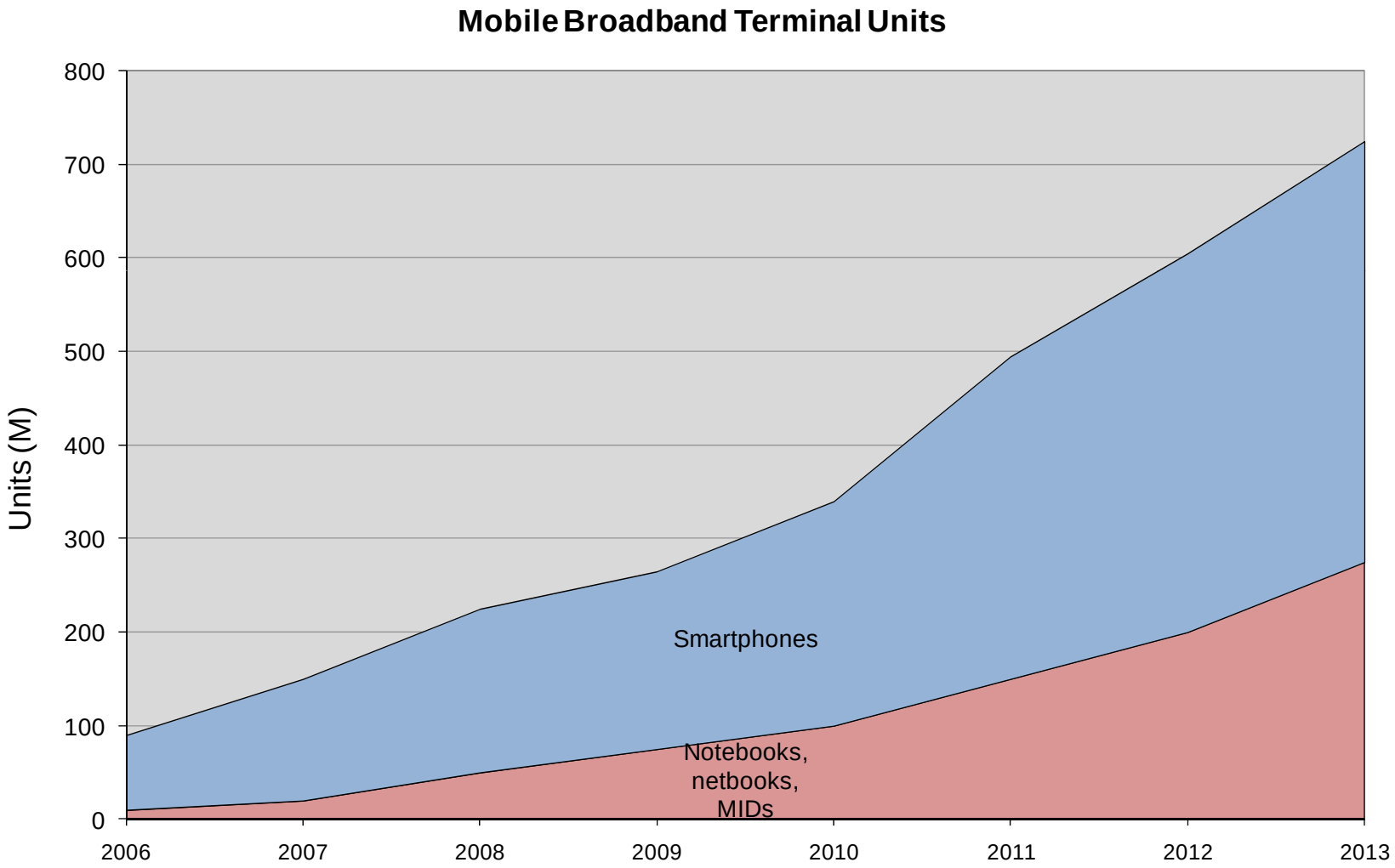
What's impressive:

- Task parallelism
- Amount of data

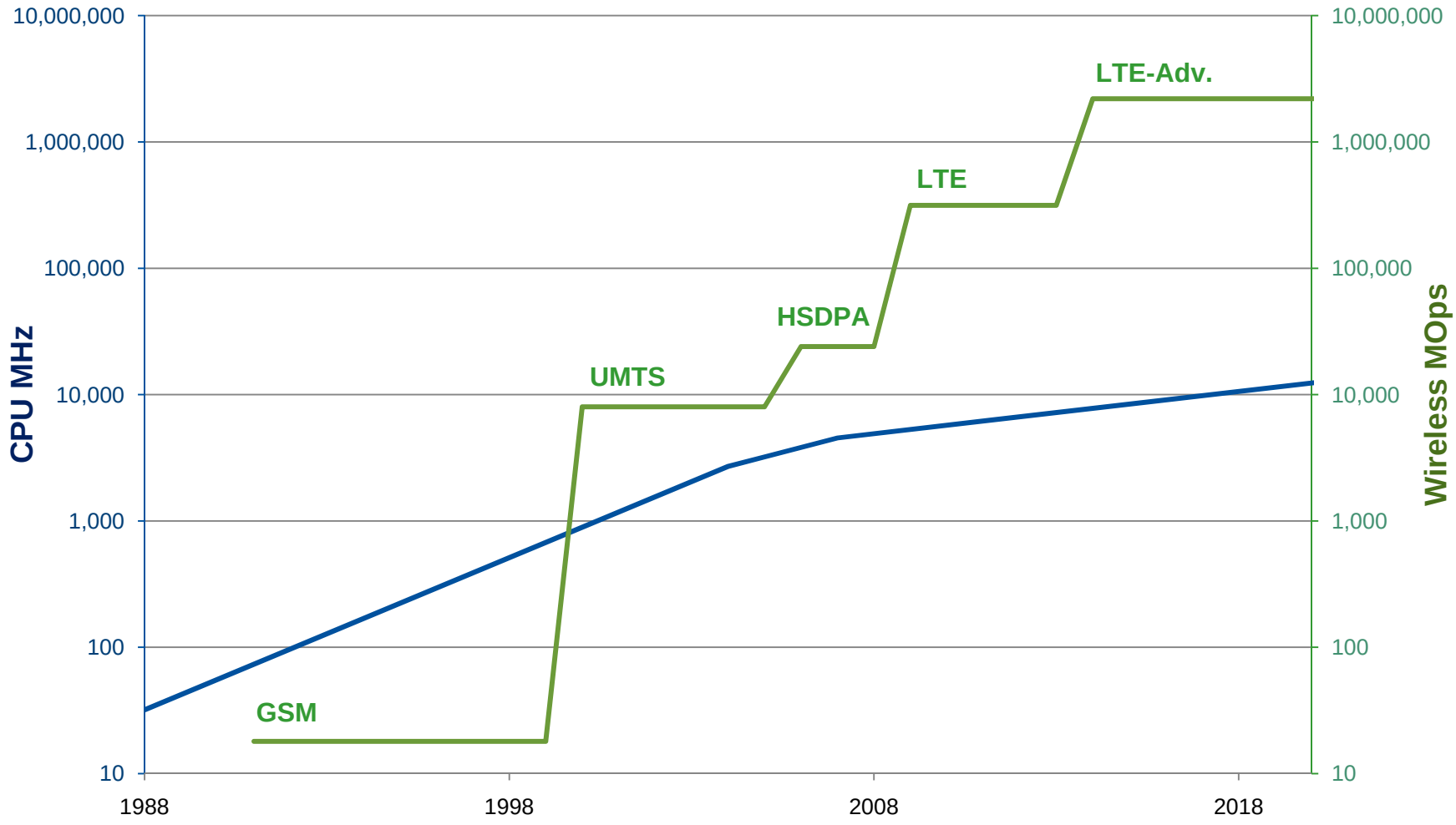
Trend #1: Silicon Density Keeps Improving
Implication: High complexity chips cost-effective



Trend #2: Mobile Internet Unit Growth
Implication: Plenty of Silicon Demand

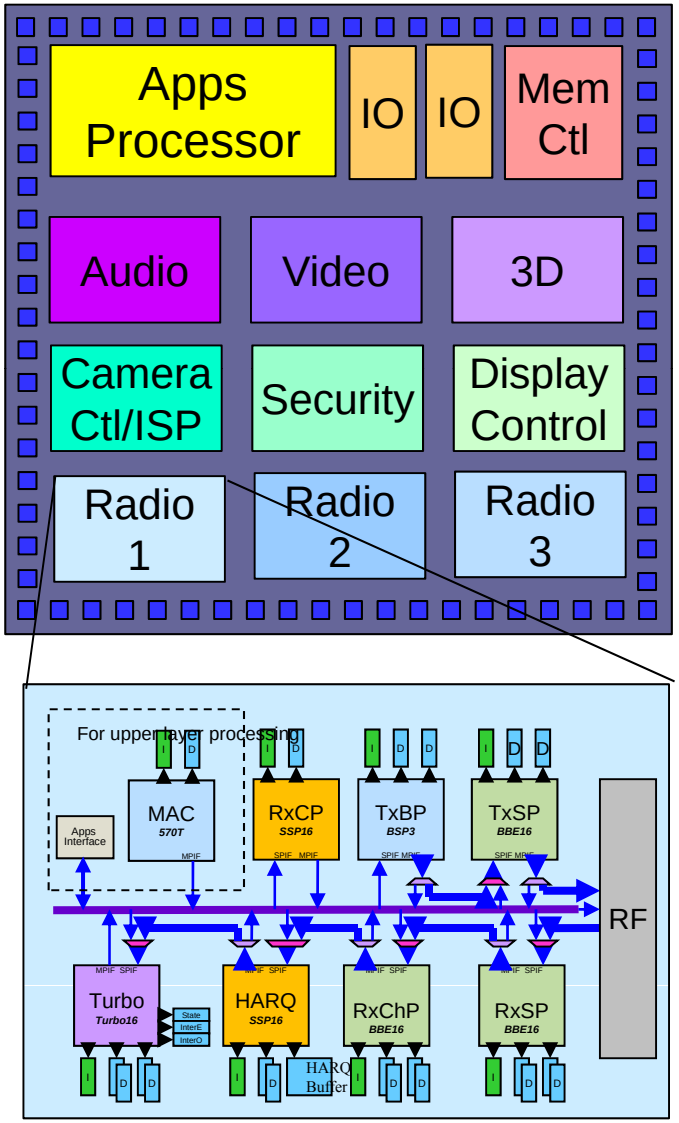


Trend #3:
Silicon Speed Can't Keep Up with Demand



CPU MHz: ITRS Roadmap 1999, 2001, 2009
Wireless Ops: Multi-Core for Mobile Phones: C.H van Berkel, ST-NXP Wireless DATE09

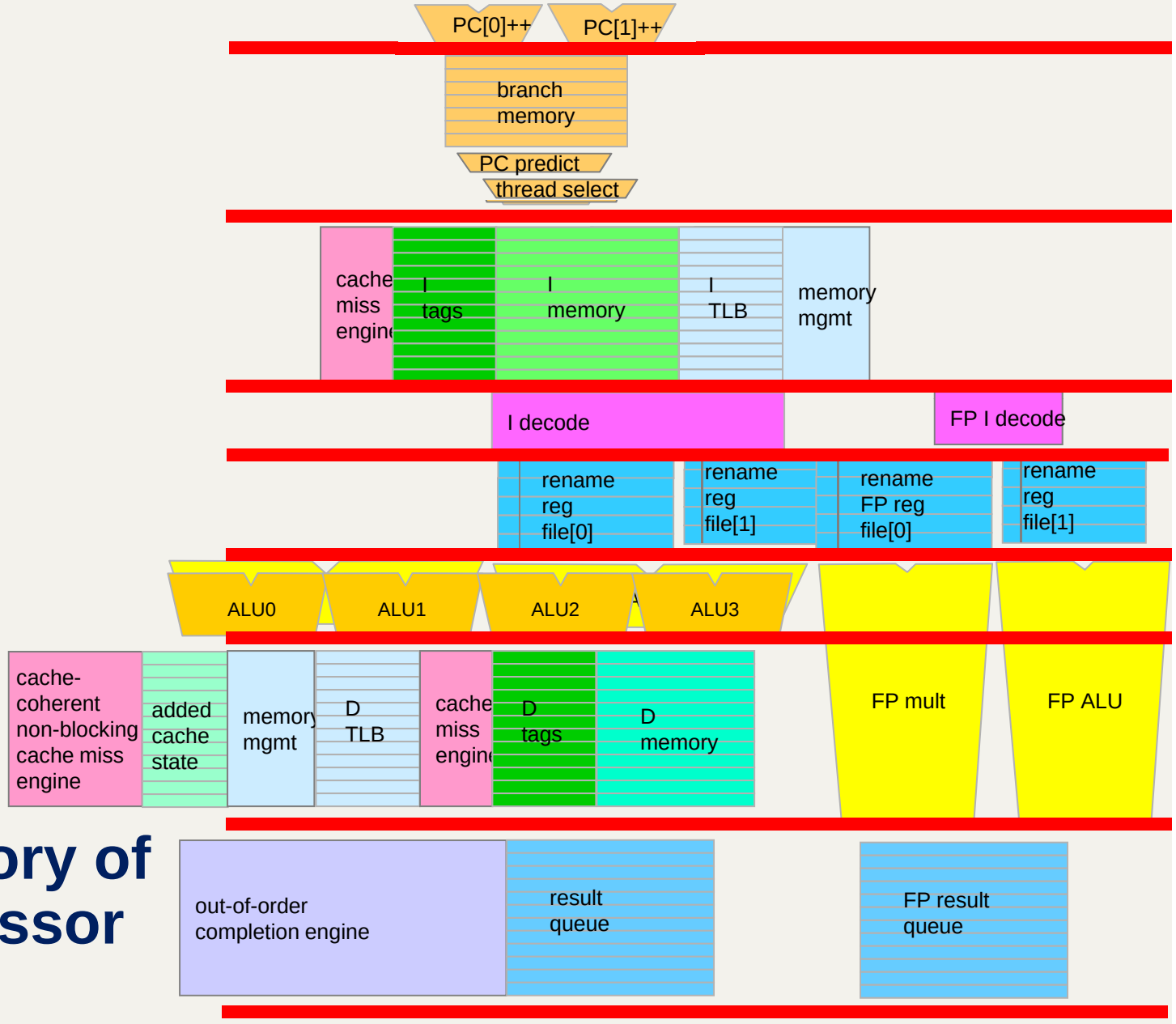
Trend #4:
System On Chip: “integrate or die”



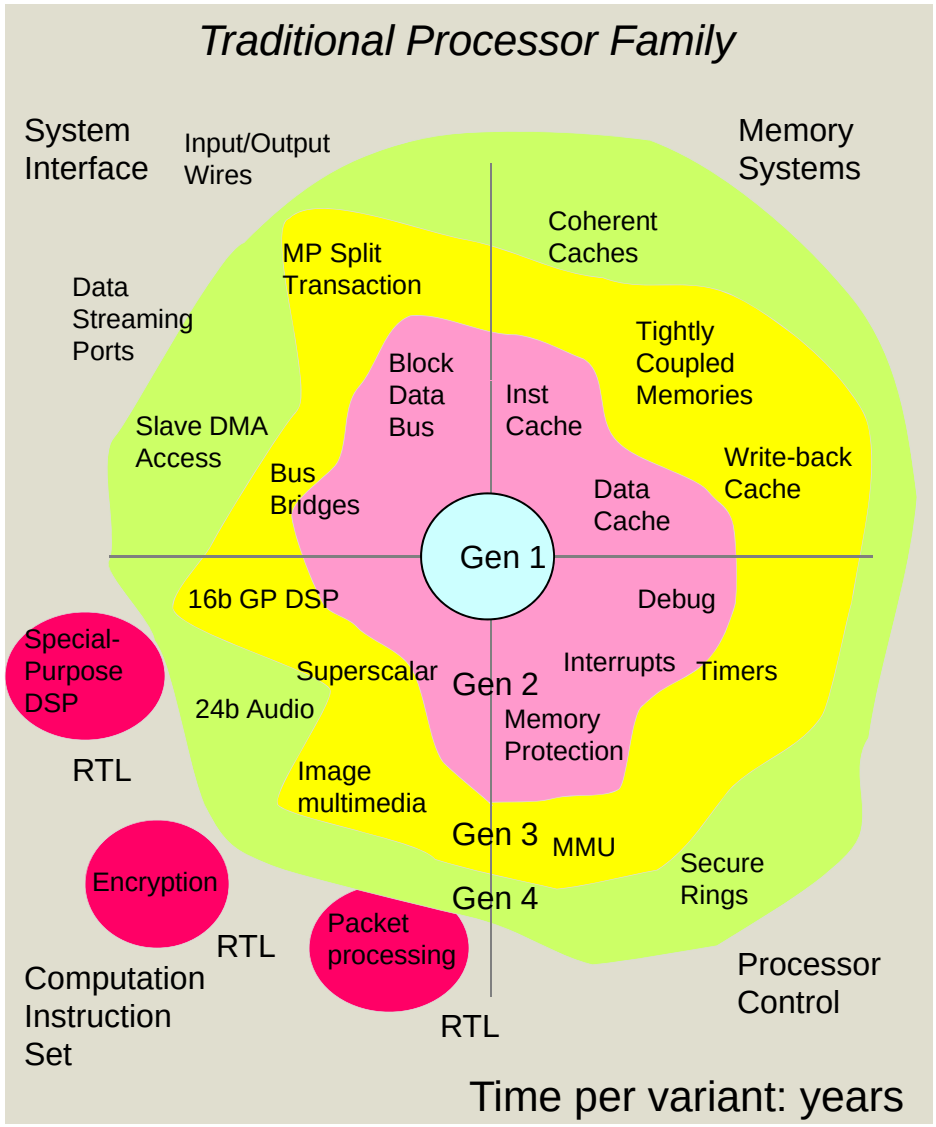
- A platform built for an application segment ➔ flexible within target domain
- An ensemble of subsystems, each:
  - developed independently
  - programmable
  - multi-core
- Wide variation in subsystems
  - fundamental data-types & ops
  - memory requirements
  - rate of change
- Requires systematic integration
  - Common programming
  - Common system modeling and verification
  - Common interconnect
  - Common performance, power estimation
  - Common VLSI flow

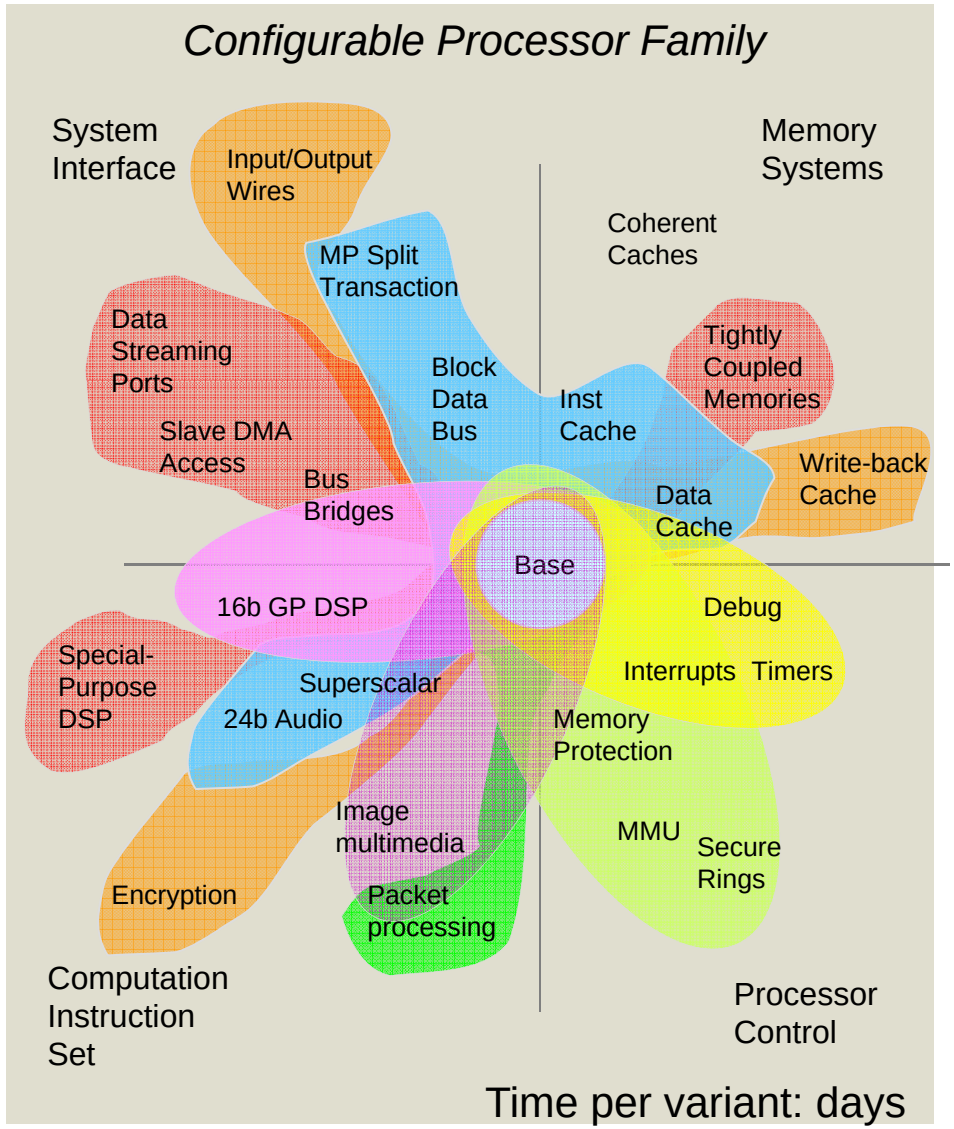
A short history of micro-processor design

- Basic micro-controller
- Micro-code
- Data width:4→8→16→32…
- General register file
- Cache and memory protect
- Pipelined load/store arch
- Floating point
- Superscalar (static or dyn.)
- SIMD multimedia ALU
- Branch prediction
- Symmetric multi-processing
- Out-of-order execution
- Simultaneous multi-thread



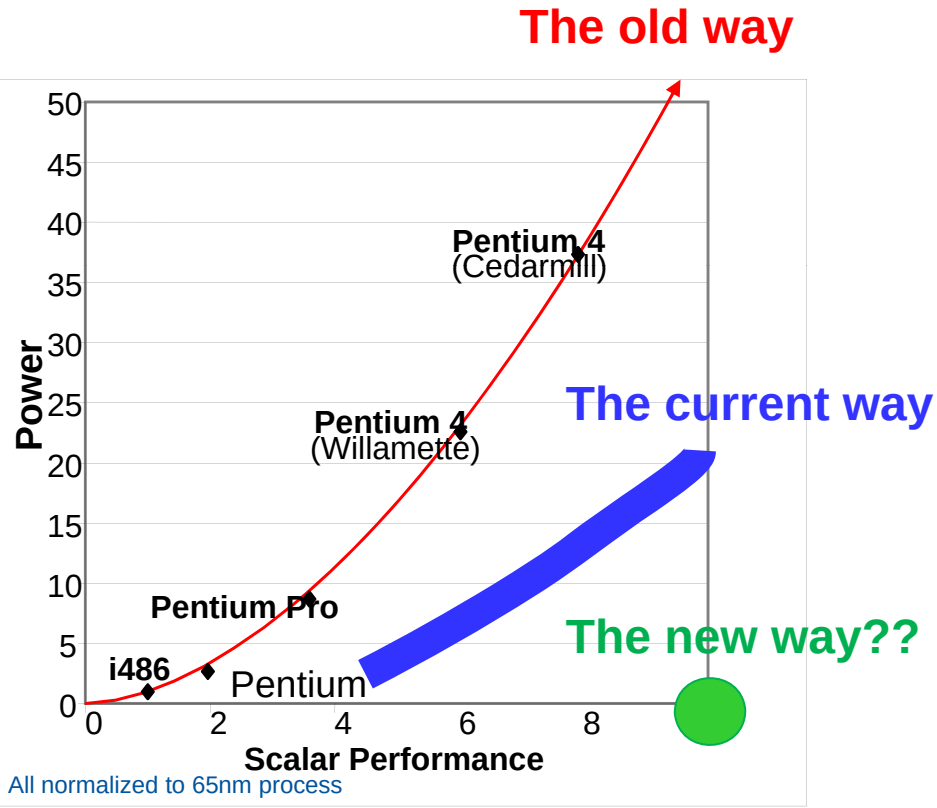
Two models of processor evolution





Where is processor power going?

Power = Performance$^{1.75}$



- Pursuit of maximum generic uni-processor performance means
  - deeper pipeline
  - minimum latency memory
  - more logic speculation

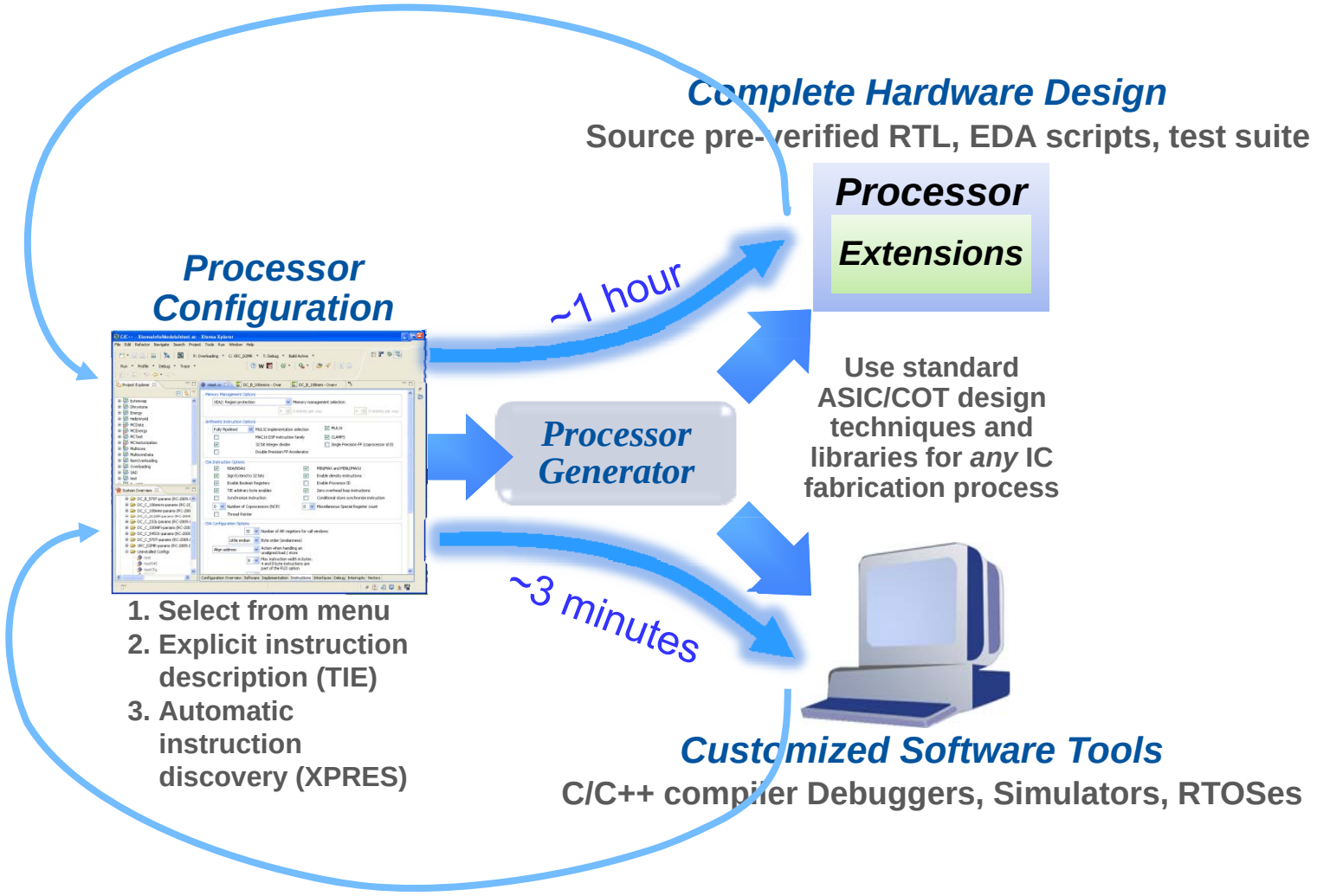
➔*More wasted energy*

- More energy-sensitive processor design has gained traction for both high-end and mobile applications
- What's the next quantum step?

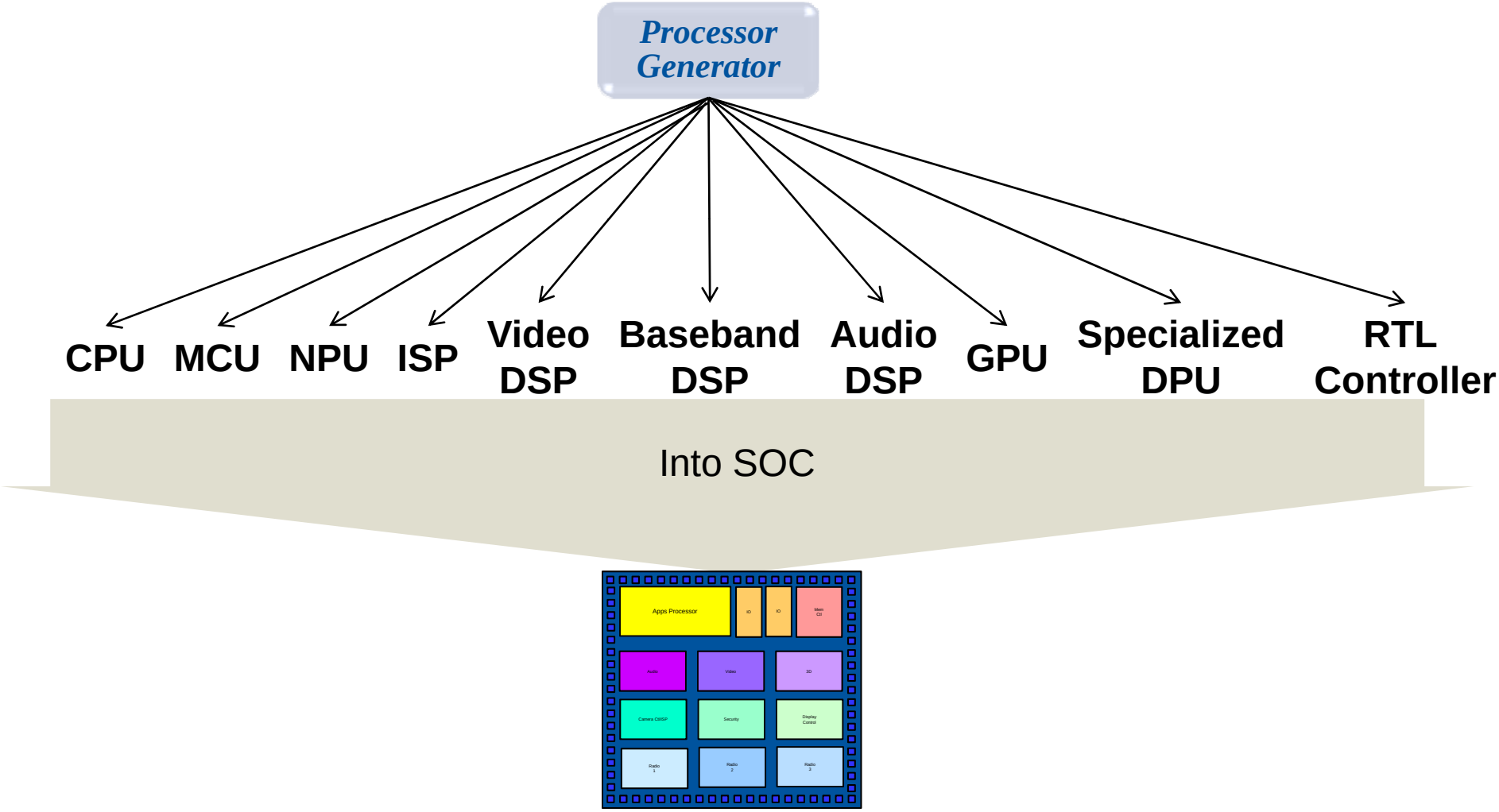
Source: John Paul Shen, Intel Microarchitecture Research Lab
WCED Panel: June 18, 2006

Processor Automation➔ Agility and Consistency

Xtensa Processor Generator

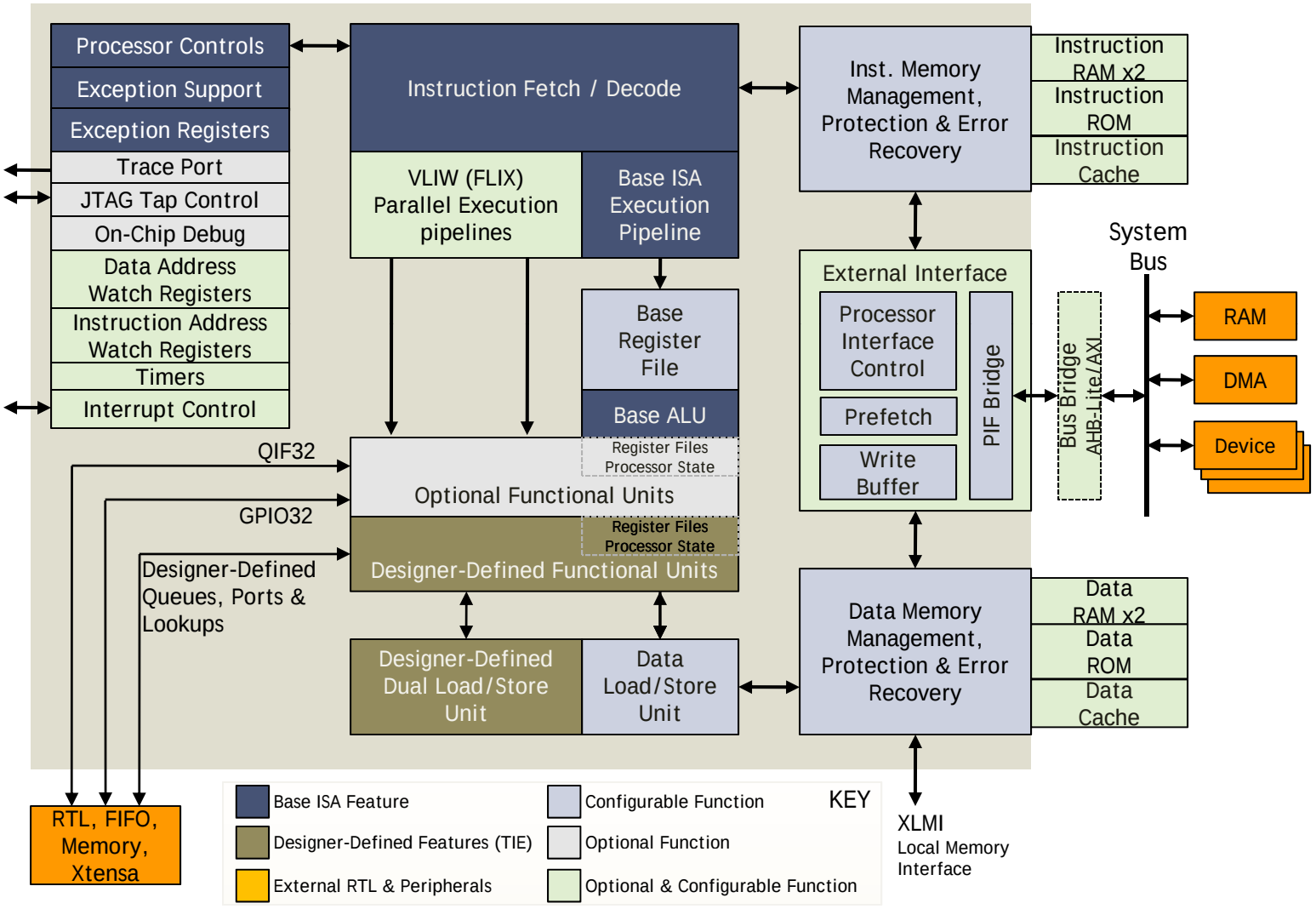


Key ingredient for “alphabet soup”



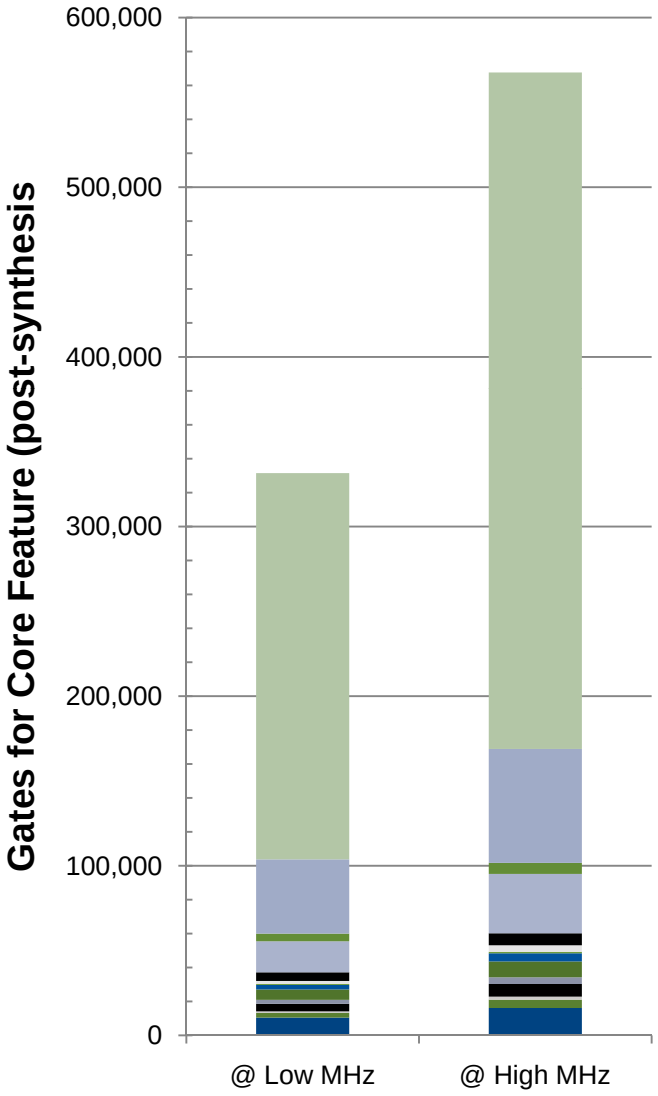
What's Inside An Extensible Processor?

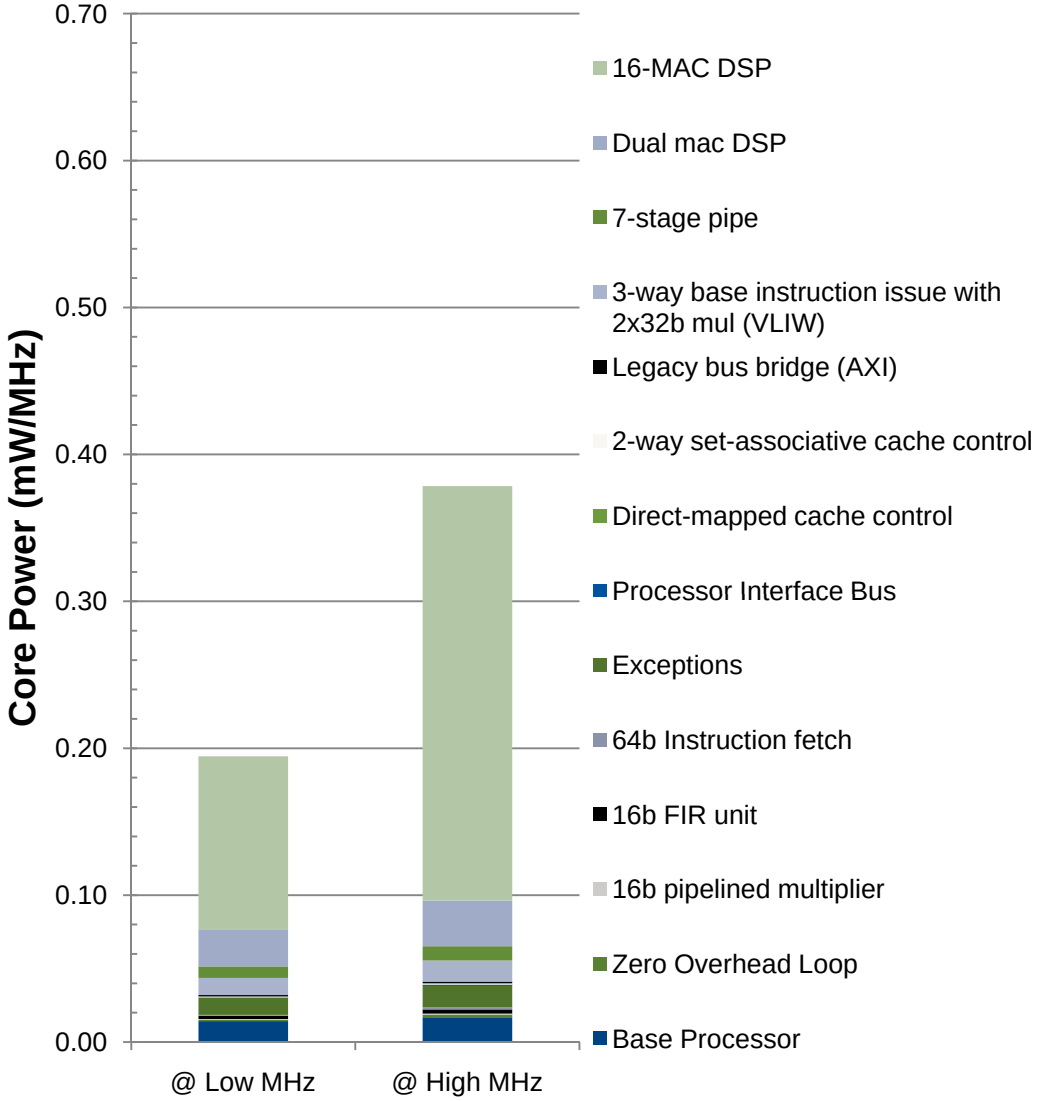
LX4 Block Diagram



The real cost of processor features

40LP process



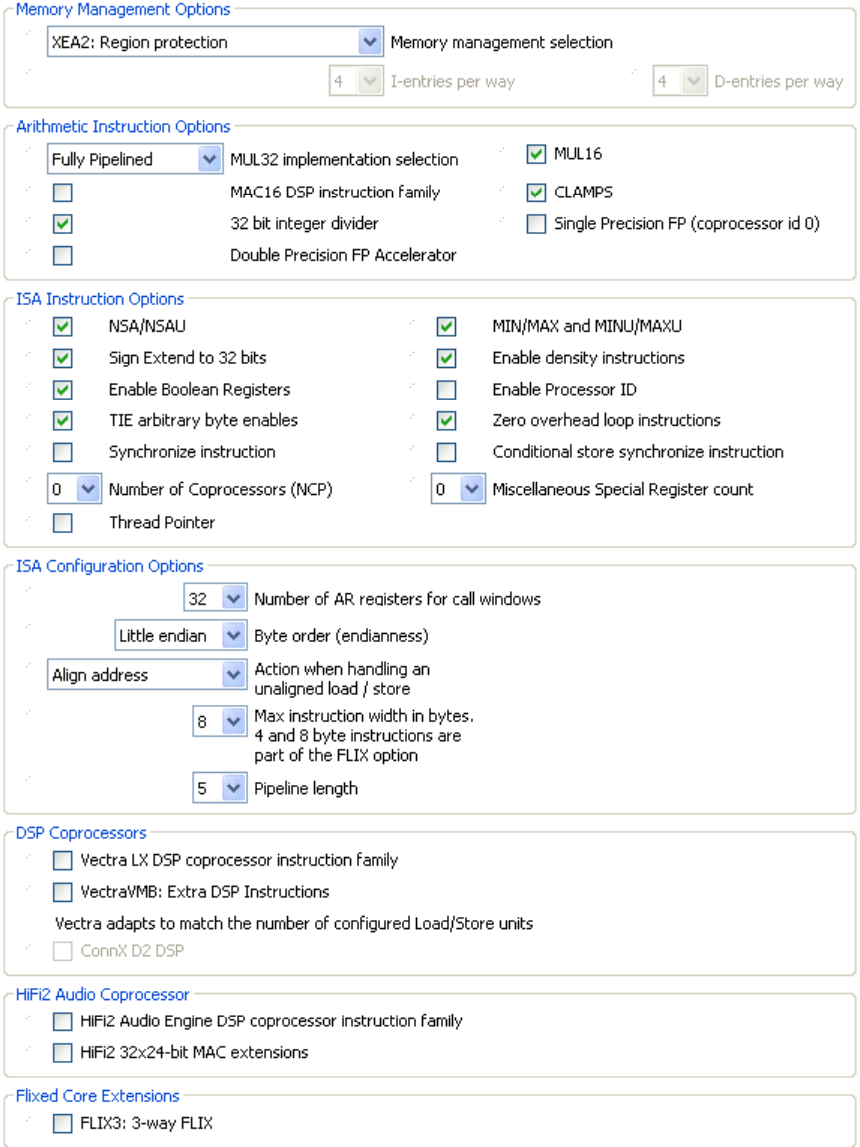


What I mean by "configurability"

Menus of options

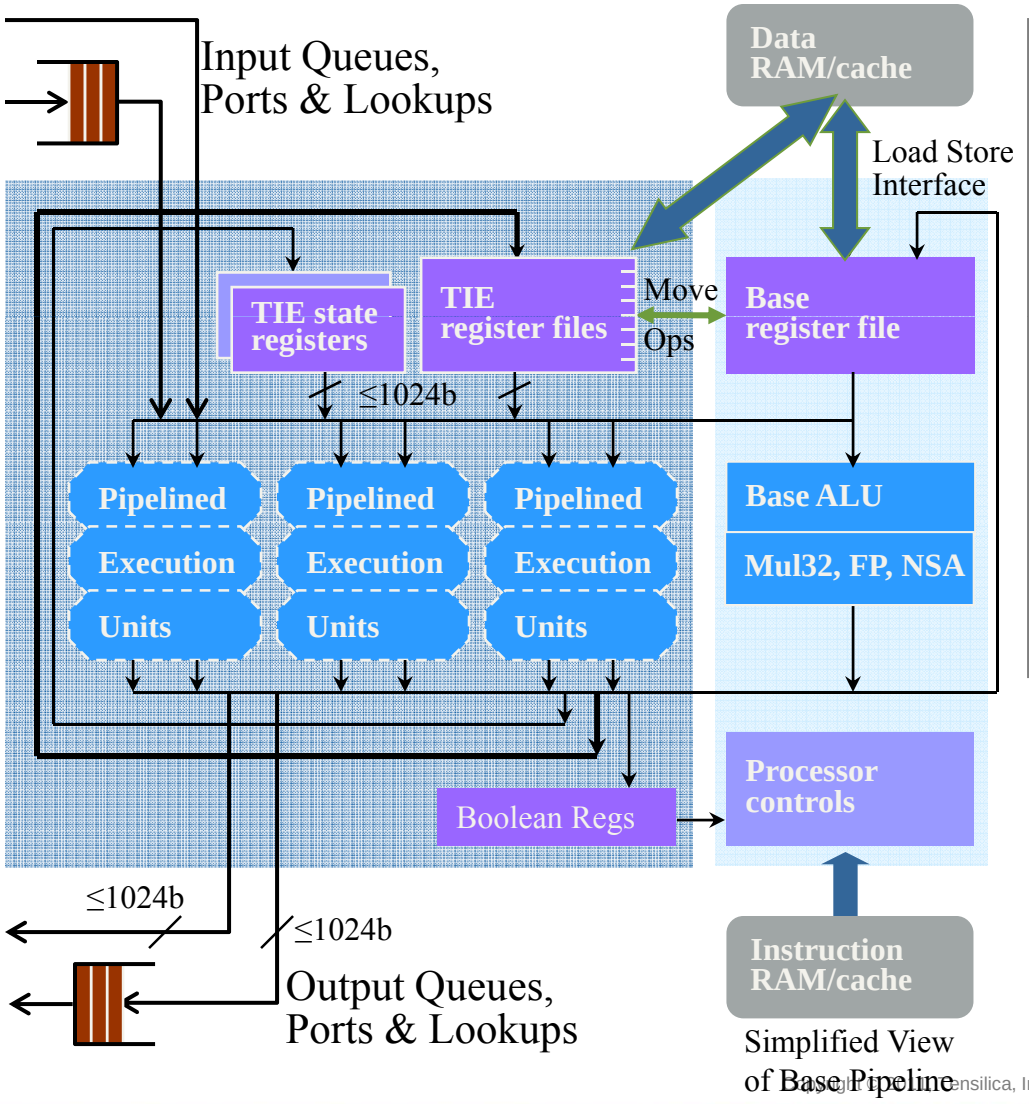
Simple menus of options

- From fine tuning of performance, power and area
  - Size, type, width and access latency of memories
  - Load/Store unit characteristics
  - Number of general purpose registers
  - Number and priority levels of interrupts
- To high-level, market-specific building blocks
  - Common functional units:
    - Floating point, multiplier, divider, NSA
  - Complex application engines:
    - HiFi2 Audio DSP
    - 16-MAC ConnX BBE16 DSP
    - Dual-MAC ConnX D2 DSP



What I mean by "extensibility"

Base processor + extension environment = optimized cores

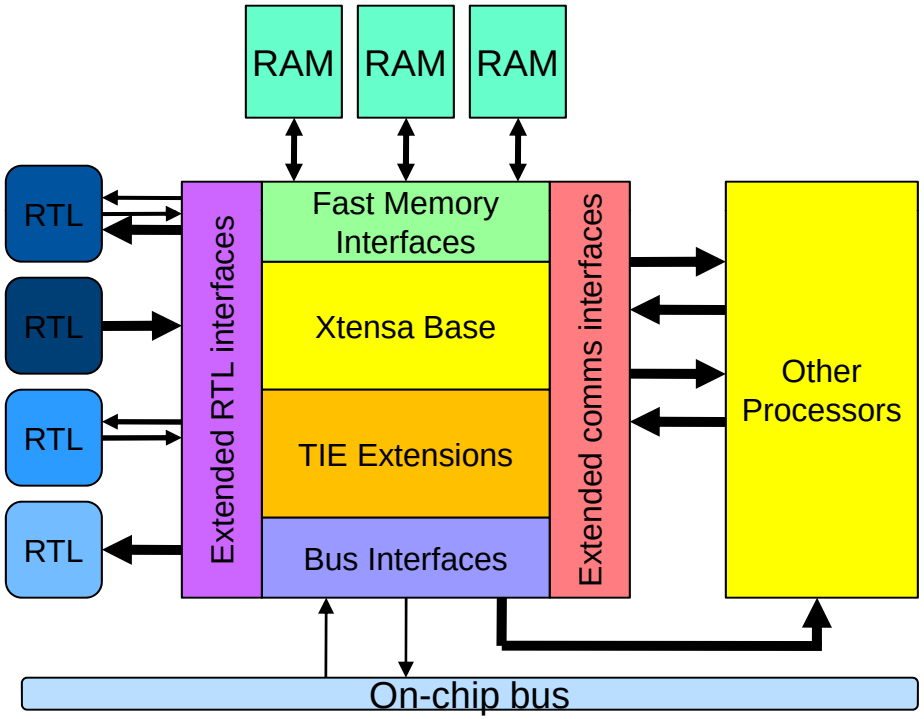


1. **Start with base Xtensa core**
2. **Add functional units from menu of config options**
3. **Add register files and state registers**
   Add corresponding new data types with automatic C/C++ compiler support
4. **Add up to 512-bit Load/Store instructions**
5. **Add multi-cycle, SIMD arithmetic and logic function units**
   Up to 64 source, destination registers
6. **Create multi-issue VLIW datapath**
7. **Add custom I/O ports, queues, and lookup interfaces**

A simple TIE example – VLIW processor with 128b data types and operations

```
regfile wd 128 16 w
operation wadd {in wd a, in wd b, out wd
 c} {} {assign c = a + b;}
format f64 64 {ls_slot, alu_slot}
slot_opcodes ls_slot {LD.wd, ST.wd,L32I,
 L16UI, L16SI, L8UI, BNE, BEQ, BEQI}
slot_opcodes alu_slot {wadd, ADD, ADDI,
 SUB,SLL,SRL,SRA,SRAI,SLLI,AND,OR, EXTUI}
```

Extensibility enables better multi-core



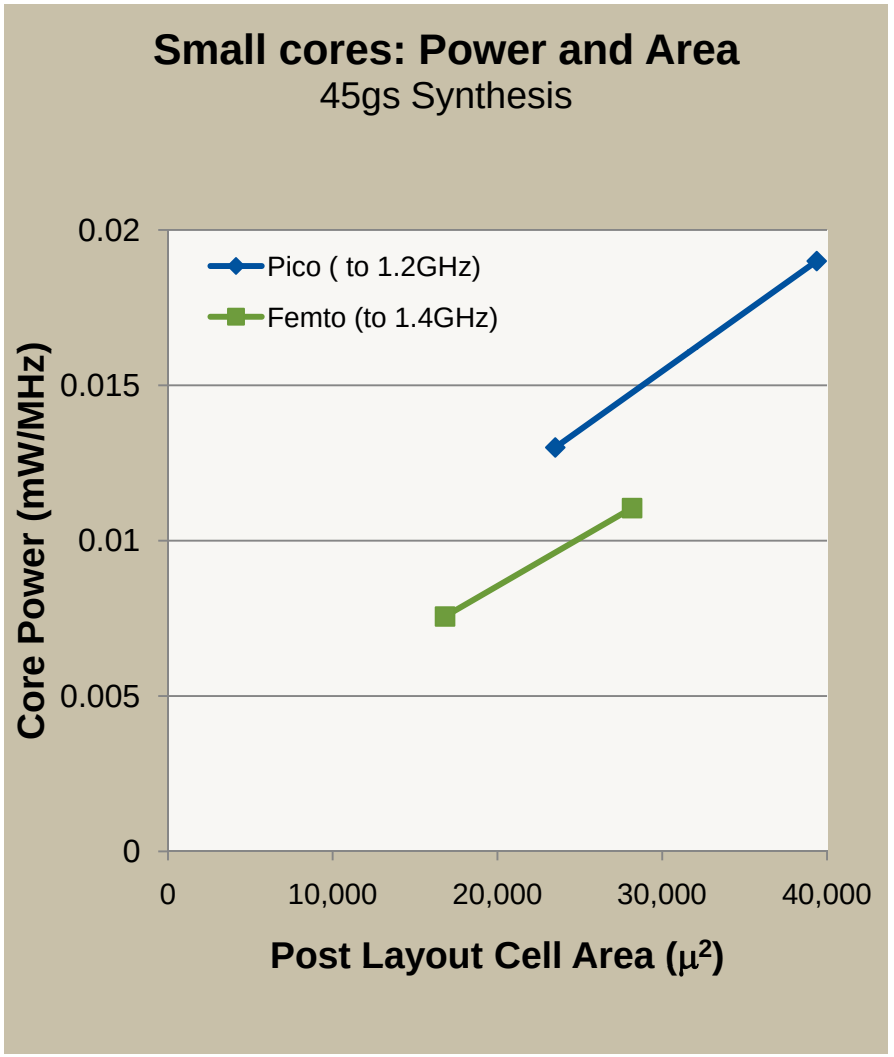
- Small base: low threshold for adding programmability
- ISA extensions: high computation efficiency
- Memory: wider and more flexible
- Standard interfaces: to buses and NOCs
- RTL interfaces: more throughput and control
- Comms interfaces: 10x mB/mW, multi-core SW libs

Pushing the envelope on small

- What happens if you really pare down the processor to essentials
  - General-purpose 32b ISA
  - Fast 5-stage pipeline
  - Local instruction/data memories
  - Optional bus interface, interrupts and exception handling, multiply/divide, zero-overhead loops
- Example Footprint (45gs):
  - Core

  + Byte-wide serial boot-loader

  + 2KB/2KB memory [yes, tiny]
  - **0.052mm²** to **0.072mm²** floorplan area
  - up to ~20 cores/mm²
  - up to ~25,000 DMIPS/mm²
- Extrapolate to 22nm process:
  - ~80 cores/mm²
  - > 120,000 DMIPS/mm²



System implications of ultra-small cores

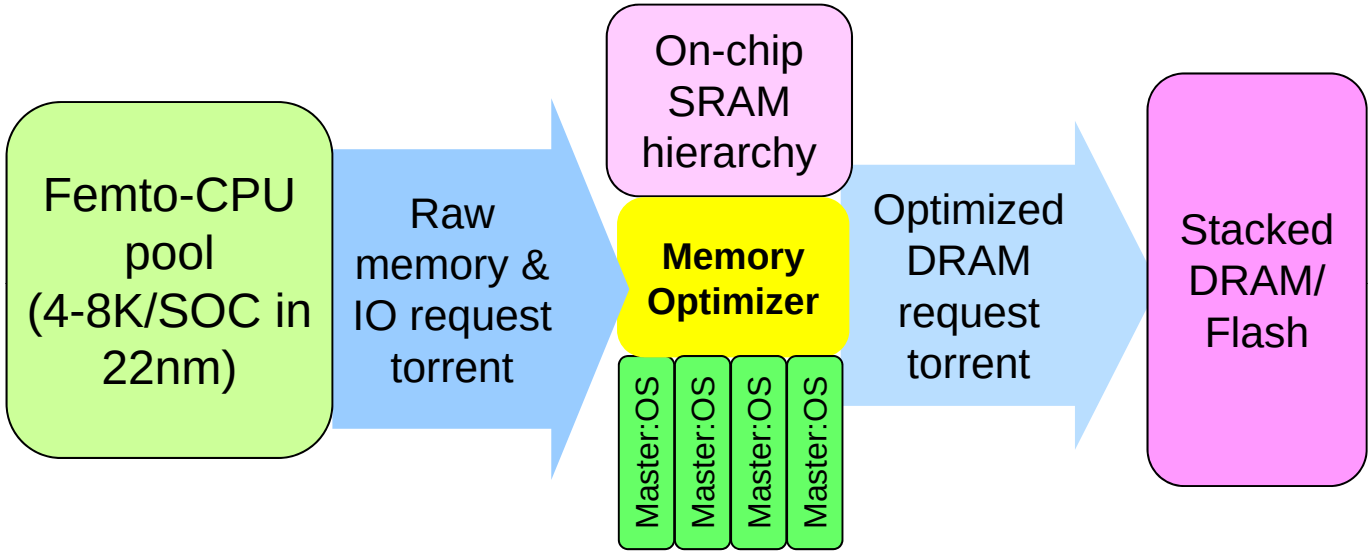
The "new NAND gate" in embedded systems:

- Distributed intelligence:
  - Dynamic QOS in on-chip communication network
  - Finite state machine control of hardwired function blocks
  - Extended for specialized computing (graphics, signal, security) – homogeneous or heterogeneous
- Extends reach of **software+processor** ecosystem to lowest layers

Radical building-block in energy-efficient server:

- Traditional server (wastes energy):
  - Few CPUs/chip: (O(10)
  - Memory optimized to minimize latency at CPU
- Minimum-energy server (wastes CPUs):
  - Very many CPUs/chip: O(1000)
  - Memory optimized to minimize energy per useful MB

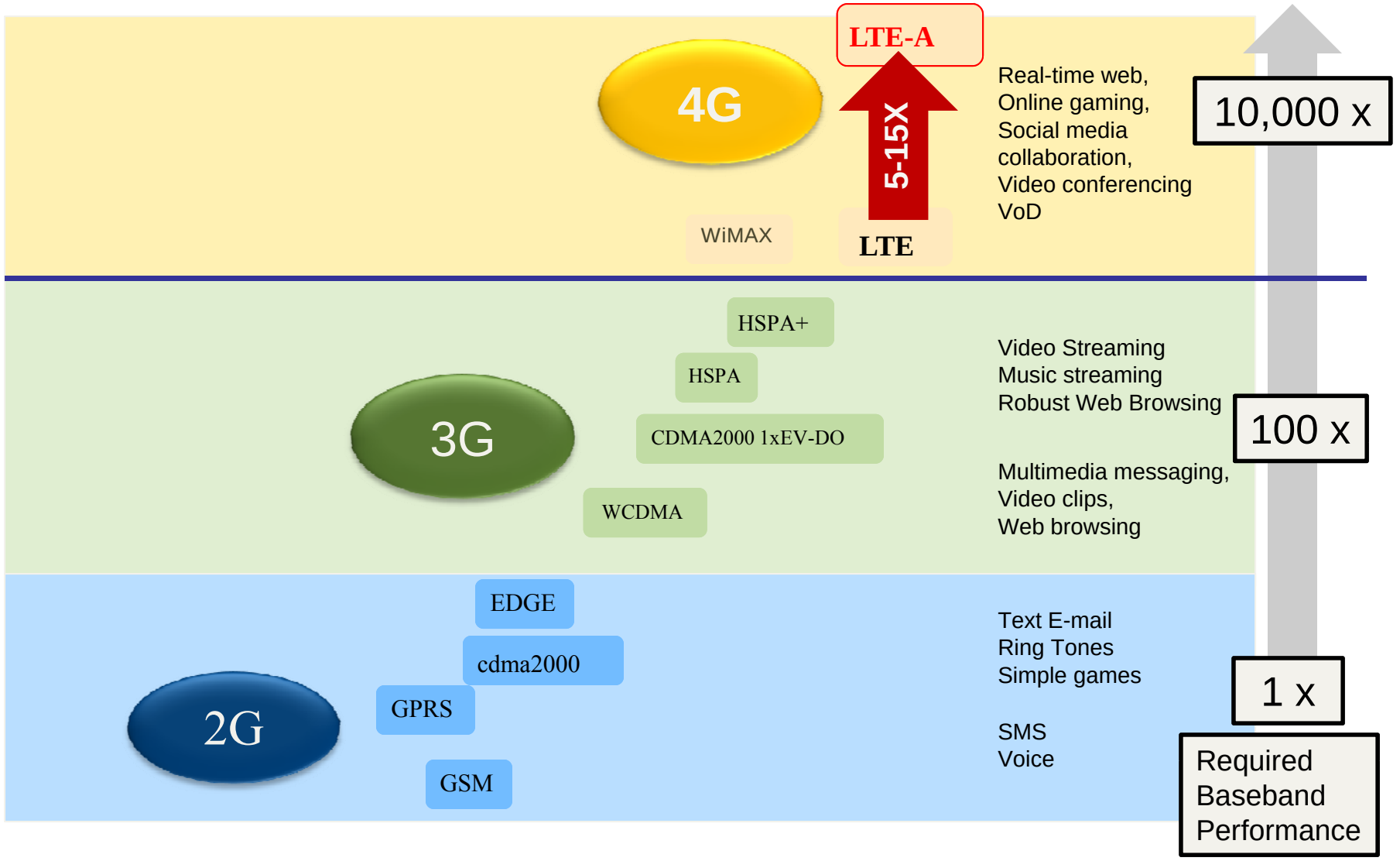
Extreme Processor #1:
Efficient service of memory torrent



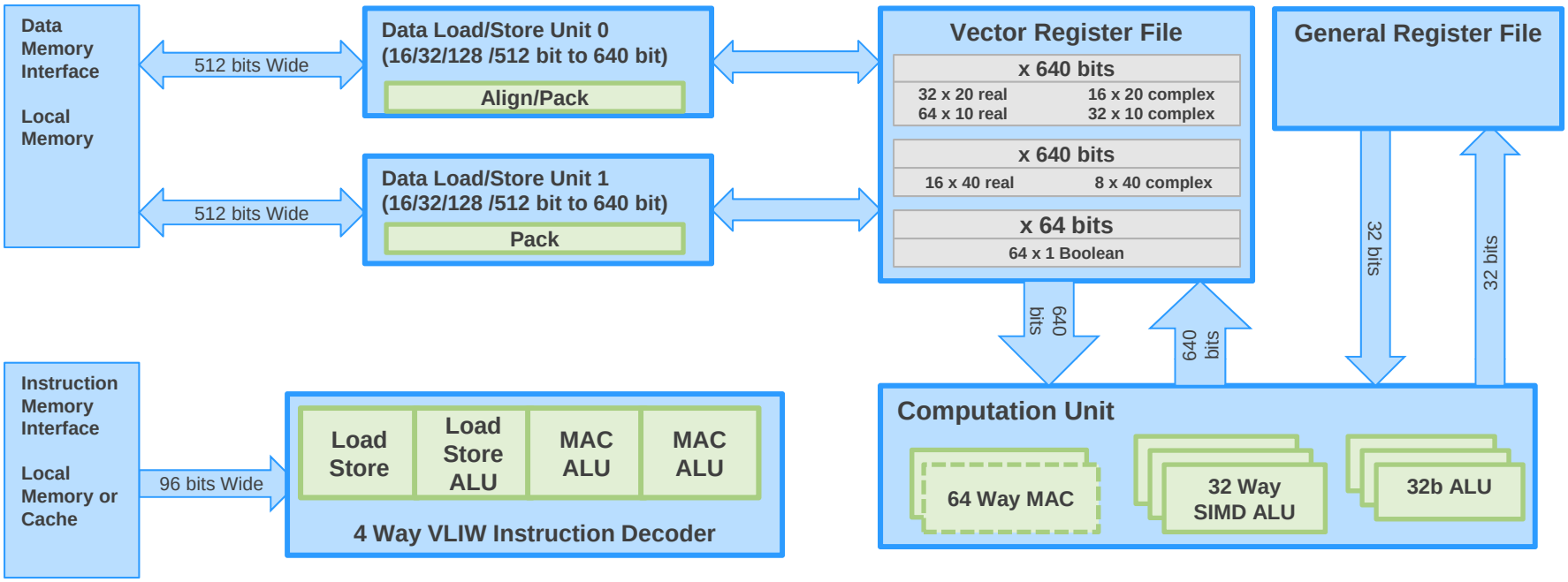
Moving from minimum latency to maximum throughput memory systems

Domain Example: 4G Wireless

Quantum Step in Required Performance



Extreme Processor #2: ConnX BBE64: >100 GMACS per second



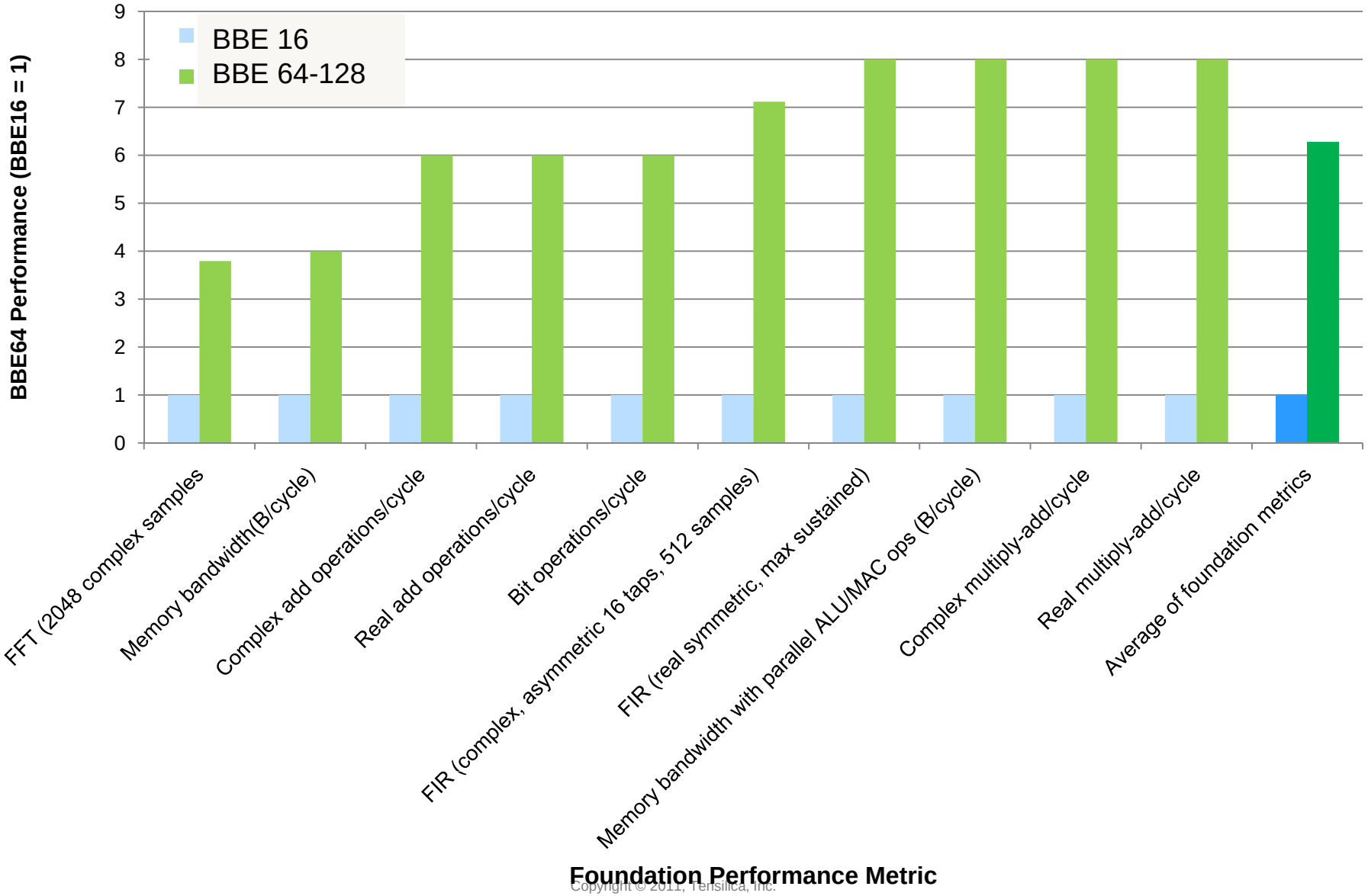
Optimized Architecture for DSP Applications

- 4-way VLIW x 32-way SIMD > 128 DSP ops/cycle
- 16/24/96b 4-issue VLIW – almost any instruction in any slot
- 128 MAC ops/cycle for matrix and filter functions
- Guard-bits on all DSP data for numerical accuracy
- Protected pipeline: interlocks/bypasses for robustness
- Support for all data types from C
  - Complex/real
  - Scalar/vector
  - Fractional/integer

High Bandwidth Configurable Memory Subsystem Interface

- Dual load/stores with dual 512b memory interfaces
- Full bandwidth on packed and unaligned data vectors
- DMA support for local data memory
- Extensible with special memory ports
- Extensible with direct connect data queues: 4 x 640b per cycle

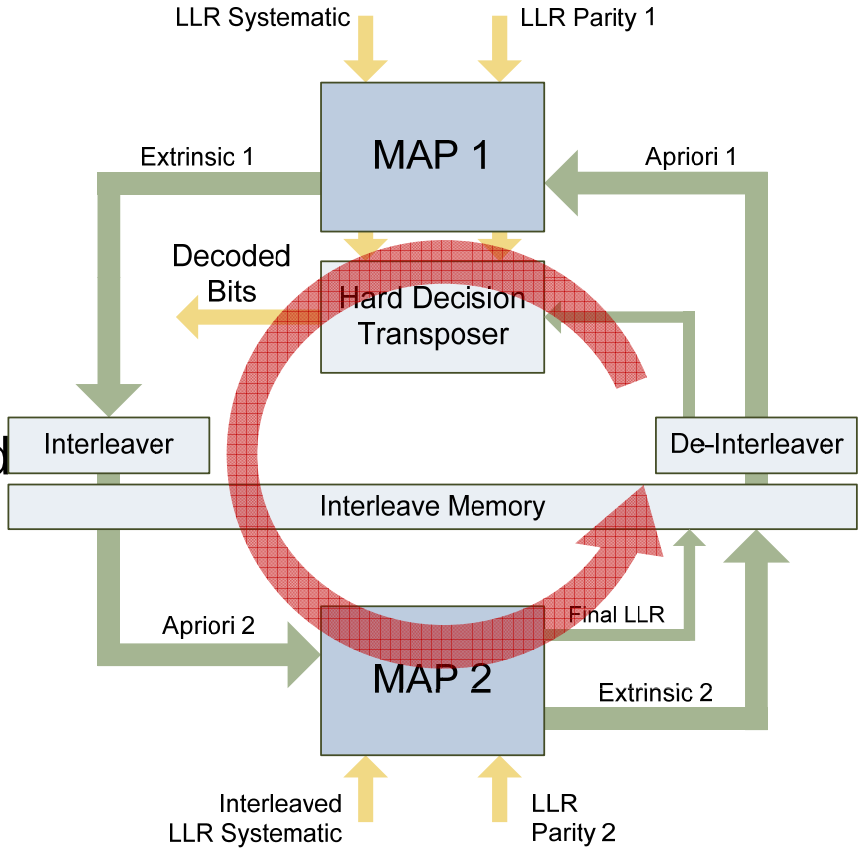
About 6x gain over current leading DSP



Extreme Processor #3

Inside a programmable Turbo decoder

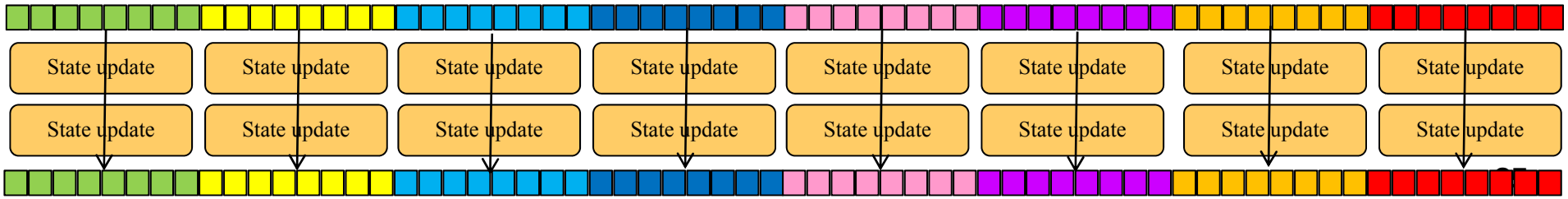
- Turbo-decoding: iterative trellis decoding for near-Shannon-limit channel coding
- Rate 1/3 coding: three soft-bit streams in (data, parityA, parityB) and one hard bit stream out
- 4-8 iterations of Turbo to reach convergence
- Up to 22 memory references per cycle (HSPA+)



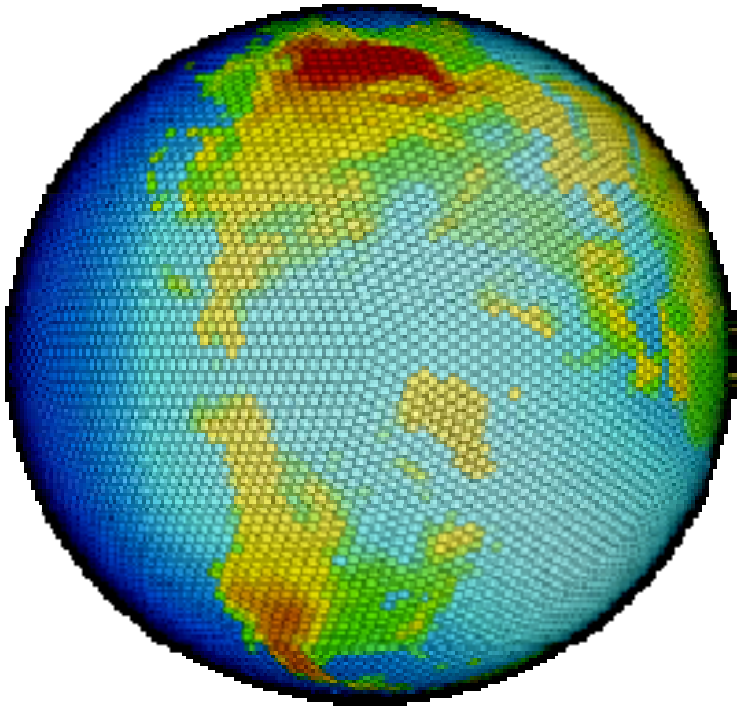
Inside the Turbo16 Processor

Huge Computation Load: >1 TeraOps

- Per state update:
  - ~25 RISC operations per state per bit X 8 states = ~200 ops per bit
- 8-parallel windows based decoding
  - Interleaving and de-interleaving operations integrated with load/ store operations to odd/even Data RAM
- 2 bits processed sequential per cycle:
  - X 2 bits per cycle = ~3200 ops per cycle (16 bits)
  - Also cuts state memory storage in half
- 150Mbps at 8 iterations requires ~380MHz: x 380MHz = 1.2 x 10^{12} operations per second
- Software-based early termination ➔lower power



Extreme Processor #4
Energy-Efficient Supercomputing



Lawrence Berkeley Lab Climate Modeling System

Lenny Oliker and Michael Wehner of Lawrence Berkeley Lab target: a much more parallel climate model

- 1.5km grid for Earth
- 20,000,000 domains @ 500 MFLOPS + 500 MB/s per domain + 20MB/s 2D mesh communications
- Complex algorithms require general-purpose programmability in double precision floating point

Technical/ Economic Challenges

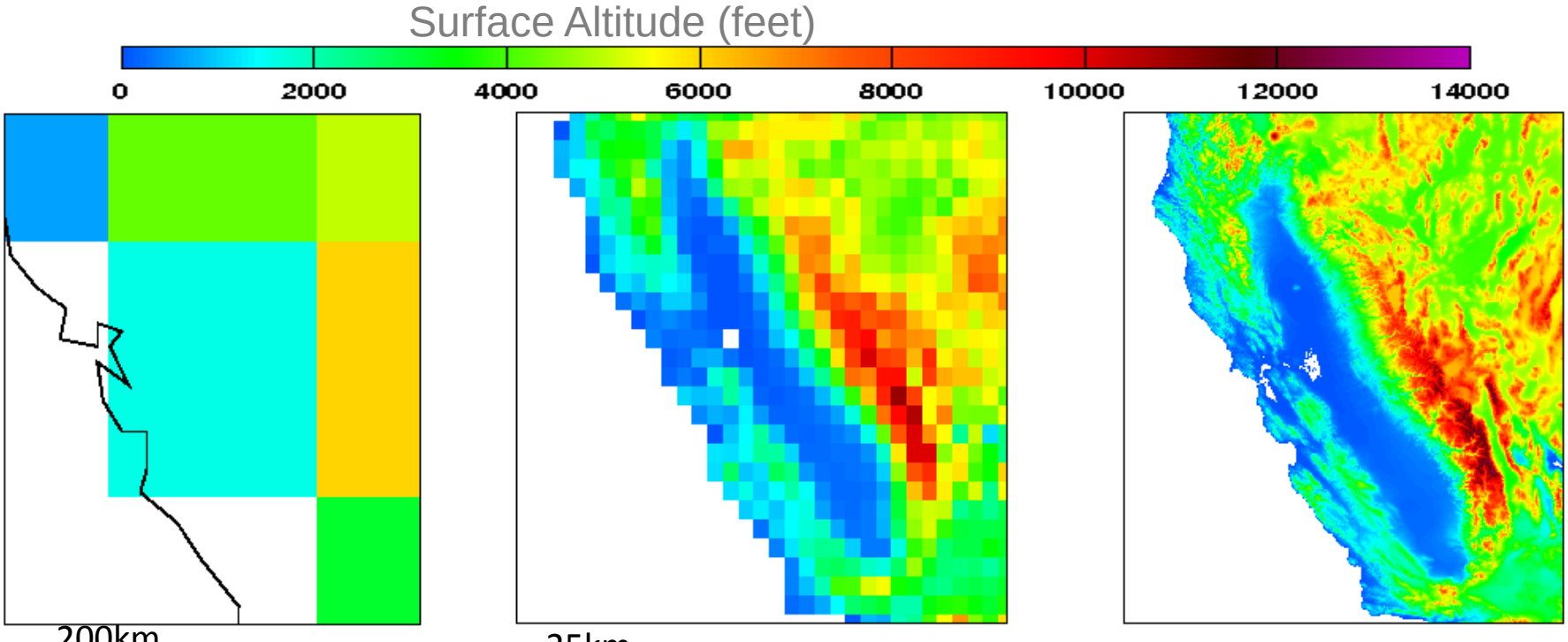
1. Variance in data-reference/ communication patterns for codes
2. Potential for extreme scalability via large-scale processing arrays
3. Size, capital and operating costs strongly correlated to system power dissipation
4. General-purpose CPUs optimized for integer applications – unimpressive performance per $, per watt

See: http://www.lbl.gov/Science-Articles/Archive/NE-climate-predictions.html

Climate Modeling System

Global Cloud System Models: A Transformational Change

- Cloud formation is critical effect in climate
- Model accuracy goes up dramatically at km-scale grid
- Higher resolution model requires 10-100 peta-flop systems



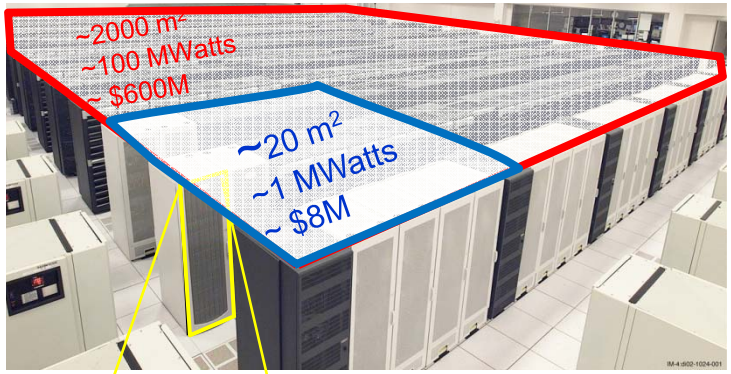
200km
Typical resolution of IPCC AR4 models

25km
Upper limit of climate models with cloud parameterizations

1km
Cloud system resolving models

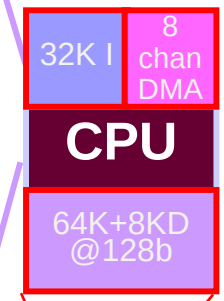
LBNL "Green Flash" Architecture: 10 PetaFLOPS System @ 2.5M CPUs ➔ 250M CPUs

ExaFLOP almost in reach

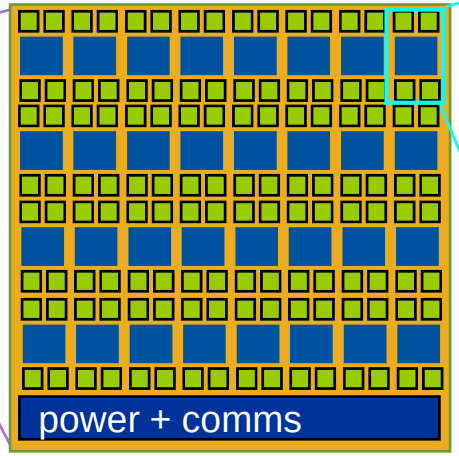


VLIW CPU:

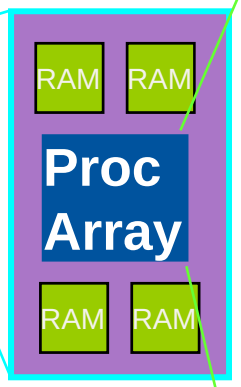
- 128b load-store + 2 DP MUL/ADD + integer op/ DMA per cycle:
- Synthesizable at 1GHz in commodity 28nm
- 0.25mm² core, 0.7mm² with inst cache, data cache data RAM, DMA interface, 0.15mW/MHz
- Double precision SIMD FP : 4 ops/cycle (4 GFLOPs)
- Vectorizing compiler, lightweight communications library, cycle-accurate simulator, debugger GUI
- 8 channel DMA for streaming from on/off chip DRAM
- Nearest neighbor 2D communications grid



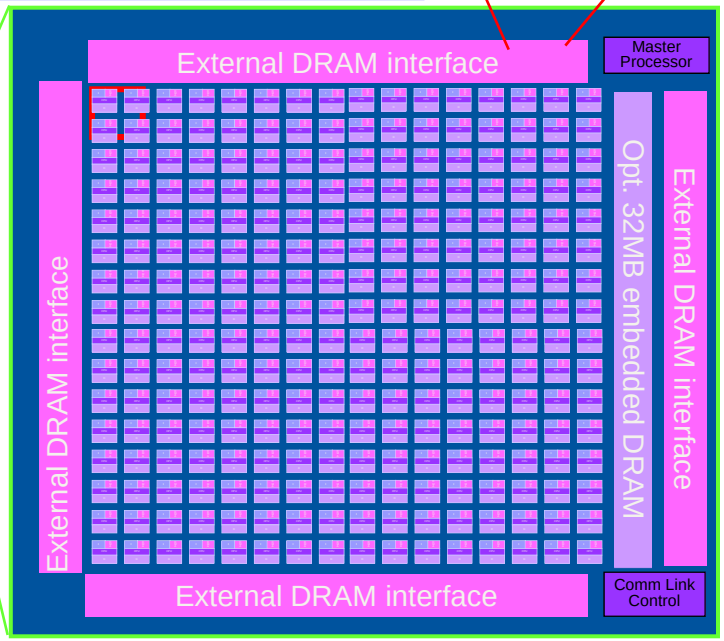
19 racks @ ~55KW



32 chip+memory clusters per board (32 TFLOPS @ 2900W



8 DRAM per processor chip: 100 GB/s



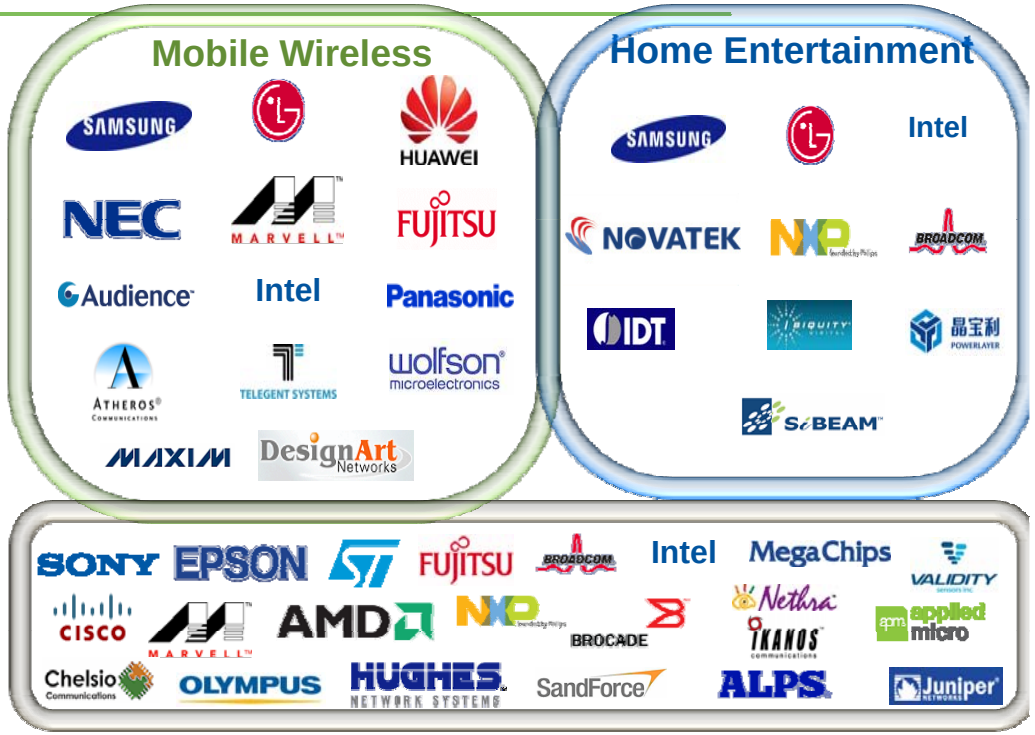
256 processors per 28nm chip >1 TeraFLOPs @ 60W

Wrap up:
Looking at long-term silicon and system trends

1. **Continued fixation on energy for mobility and cost:**
   - memory access cost >> computing cost
   - mW per [useful] MBps most relevant metric
2. **Parallelism in cloud >> parallelism in terminal devices:**
   - massive ~homogeneous servers take increasing fraction of compute
   - terminal devices dominated by access (radios) and presentation (media)
3. **Expertise in data-intensive real-time functions essential to leading in volume markets:**
   - wireless/DSP
   - multimedia: audio, video, imaging/recognition, graphic/rendering
4. **"Sea-of-processors" SOC design increasingly real:**
   - design simplicity with processor generation
   - greater technical and market flexibility

Warning: Advertisement
Tensilica At A Glance

- **Business Model:** Licensing of "Dataplane Processor Unit" IP to semiconductor and system OEM SoC design teams
- **Customers:** 160+ companies, including 2/3 of top semiconductor manufacturers
- **Focus Markets:** Mobile Wireless and Home Entertainment



- **Products:**
  - *ConnX Baseband Engines for Wireless: BBE16, BBE64, D2, SSP16, BSP3, Turbo16 + software*
  - *HiFi Audio DSPs: 330HiFi, HiFi2, HiFi2 EP*
  - *Xtensa Processor, Processor Generator and Software Environment*
- **Volume:** >500M cores/year, fastest architecture to 1B cores
- **Company status:** Largest privately-held IP firm, profitable