Lecture 19: Architecture, Arithmetic & Technology: some thoughts on the road ahead

M. J. Flynn

Semiconductor Industry Roadmap

Semiconductor Technology Roadmap (2001)

Year	**2001**	**2004**	**2007**	**2013**
Technology generation (nm)	130	90	65	32
Wafer size (mm)	300	300	300	450
Defect density (per m^2)	1356	1356	1356	1116
µP die size (mm^2)	310	310	310	310
Chip Frequency (MHz)	1767	4000	6700	19350
MTx per Chip (Microprocessor)	276	552	1204	4424
MaxPwr(W) High Performance	130	160	190	251

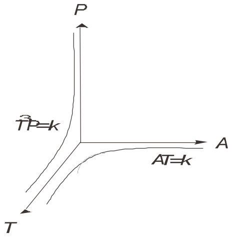
What this means to Computer Architecture

Tradeoffs in time (performance), power and area (cost). (TPA product)

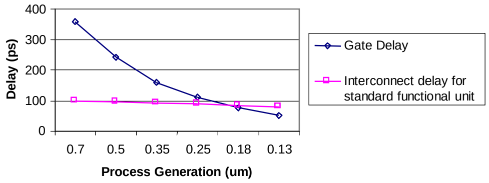
Message: System not Processor Architecture

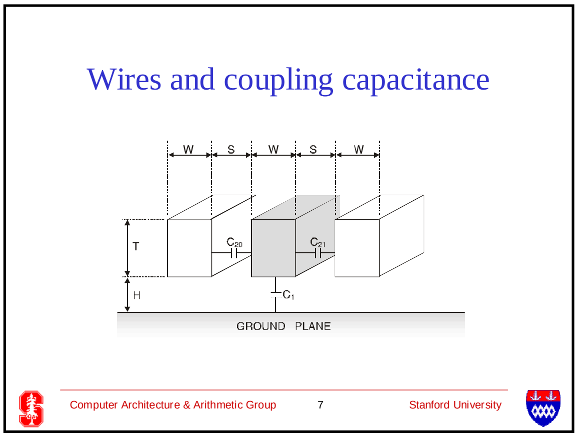
- Power not clock rate
- System not processor integration
- Integration of what?
  - Wireless and optical interconnections
  - Voice and sensor I/O
  - Crypto, Video, Audio, etc.
  - Multi user displays, keyboards, sensors, etc.

Performance: Time, Area and Power Tradeoffs

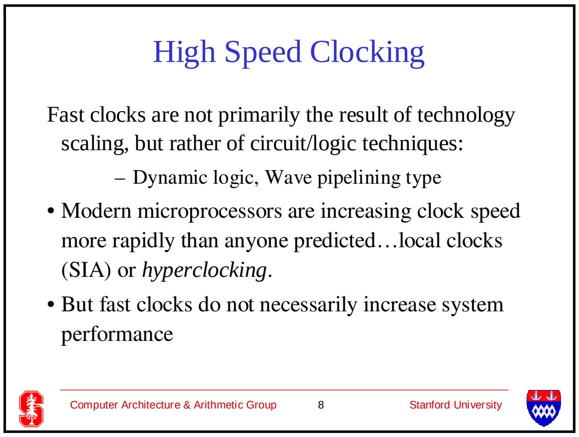


Performance lesson: wires not gates determine delay.





Wires and coupling capacitance

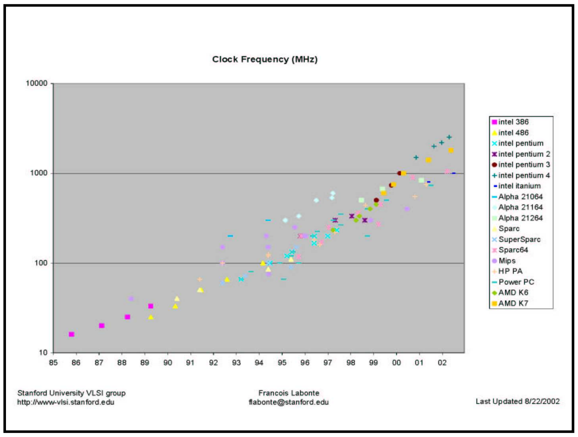


High Speed Clocking

Fast clocks are not primarily the result of technology scaling, but rather of circuit/logic techniques:

- Dynamic logic, Wave pipelining type

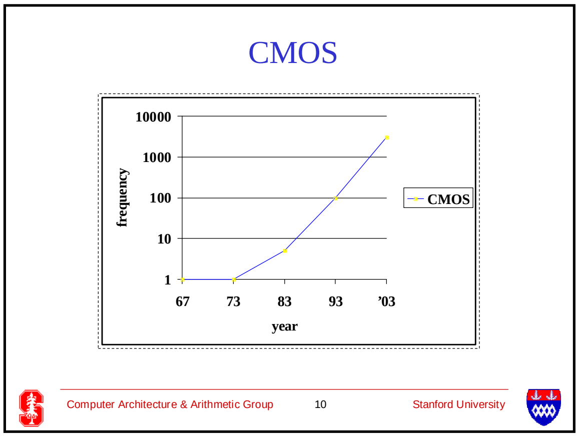
- Modern microprocessors are increasing clock speed more rapidly than anyone predicted…local clocks (SIA) or *hyperclocking*.
- But fast clocks do not necessarily increase system performance



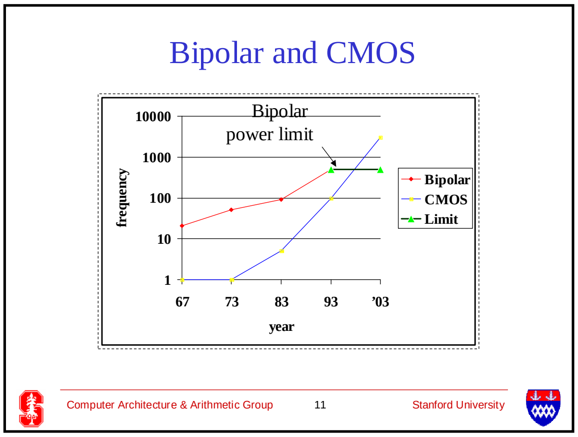
Stanford University VLSI group
http://www-vlsi.stanford.edu

Francois Labonte
flabonte@stanford.edu

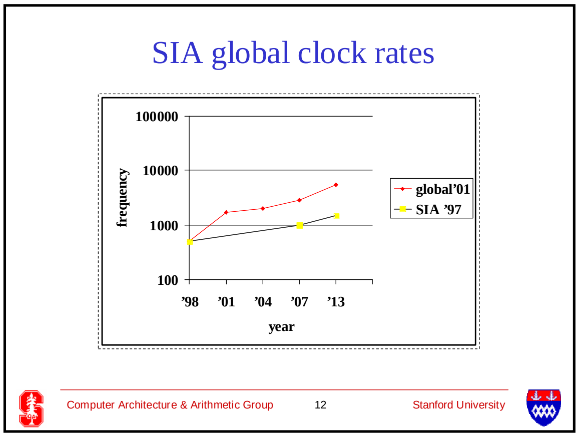
Last Updated 8/22/2002



CMOS

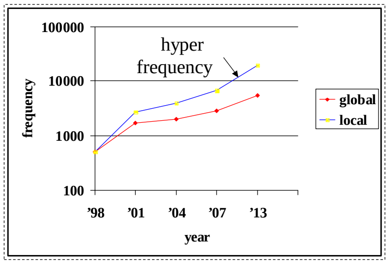


Bipolar and CMOS



SIA global clock rates

SIA local & global clock rates



What does this mean to arithmetic?

- As pipe segments decrease from say 30 to 12 or less FO4's the pressure on adder design is most noticeable…especially at 64b and m-way address adders.
- Multiply (and divide) are naturally segmented into about 6 FO4's – pp generation and reduction – and are more adaptable.

Performance and the memory "wall"

- BUT regardless of clock rate, performance is limited by the predictability & supply of instructions and data operands from memory and cache.
- The access time to memory depends on wire delay which remains constant with scaling.
- Even with significant hardware support (area) unpredicted events (missed branches, etc) force the processor to access memory. At some point limiting the performance.

Power: and the curse of frequency

$$P_{\text{total}} = \frac{C \cdot V^2 \cdot \text{freq}}{2} + I_{\text{leakage}} \cdot V + I_{\text{sc}} \cdot V$$

While V_{dd} and C decrease, frequency increases at an accelerating rate thereby increasing power density

As V_{dd} decreases so does V_{th};this increases I_{leakage} and static power

Net: while increasing frequency may not increase ***performance*** by much it certainly does increase ***power***

Power: the new frontier

- Cooled high power: >100w/ die
- High power: 20-40w/ die … plug in supply
- Low power: 0.1-2w / die.. rechargeable battery
- Very low power: 1-100mw /die ..AA size batteries
- Extremely low power: 1-100 microwatt/die and below (nano watts) .. button batteries

Battery Energy & use

type	*energy capacity*	*time*	*power*
rechargeable	10,000 mAh	50 hours (10-20% duty)	400mw-4w
2xAA	4000 mAh	½ year (10-20% duty)	1-10 mw
button	40mAh	5 years (always on)	1uw

Power is important!

By V_{dd} and device scaling $$\frac{\text{freq}_2}{\text{freq}_1} = \sqrt[3]{\frac{P_2}{P_1}}$$

- By scaling alone a 4***x*** slower implementation may need *only* 1/64 as much power.
- Gating power to functional units and other techniques should enable 100MHz processors to operate at O(10^{-1}) to O(10^{-3}) watts.
- Goal: O(10^{-6}) watts.

Extremely Low Power

System: suppose low power was the performance metric, rather than cycle time.

Very Low power (µW) can be achieved:

- Perhaps sub thresh hold circuits
- Clock rate may be limited to 10 MHz (?)
- Achieve 1 to 5 year battery life
- Lots of area, *BUT* how to get performance
  - Very efficient architecture and segmented powering.
  - Very efficient software & operating system
  - Very efficient signal processing & arithmetic
  - Very effective parallel processing

What does this mean to arithmetic?

- Power optimized arithmetic is a generally open topic.
- Minimize state transitions, clocking.
- Scale devices: path delays $P_{max}= P_{min}$
- Optimize precision

Area: scaling silicon

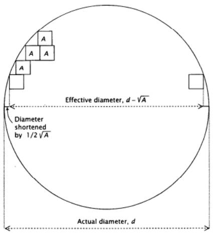
Wafer size + die size = cost

Feature size + die size = performance

The basics of wafer and die size

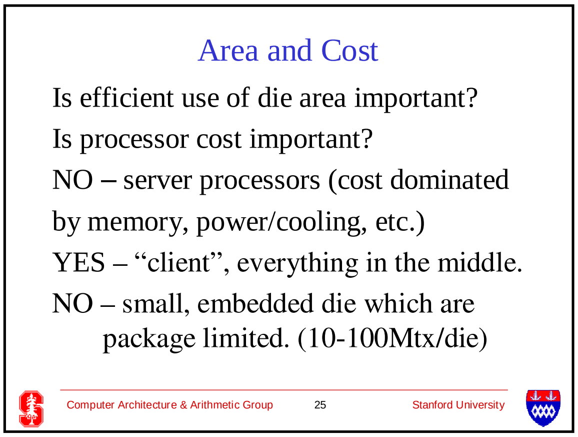
- Wafers, 30 cm or so in diameter produce ($\pi d^2/4$)/A dice. Die area, A, is 0.5- 2cm^2 about 700 1cm^2 dice per wafer.
- Yield (% of good die) is $e^{-\rho A}$ where ρ is the defect density, @ ρA=1; Yield is 37%

ρA =.2; Yield is 80%


A
A
A
A
Effective diameter, $d - \sqrt{A}$
Diameter shortened by $1/2\sqrt{A}$
Actual diameter, d

The basics of wafer fab

- A 30 cm, state of the art (ρ = 0.2-.5) wafer fab facility might cost \$3B and require \$5B pa to be profitable…that's at least 5M wafer starts and almost 5B 1cm die /per year. At O(\$1000) per wafer that's \$1/die with, say a die (f=0.1u) has 100-500 M tx /cm^2.
- So how to use them?



Area and Cost

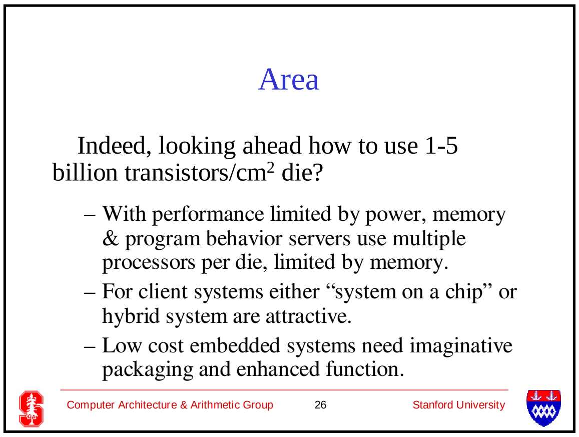
Is efficient use of die area important?

Is processor cost important?

NO – server processors (cost dominated by memory, power/cooling, etc.)

YES – “client”, everything in the middle.

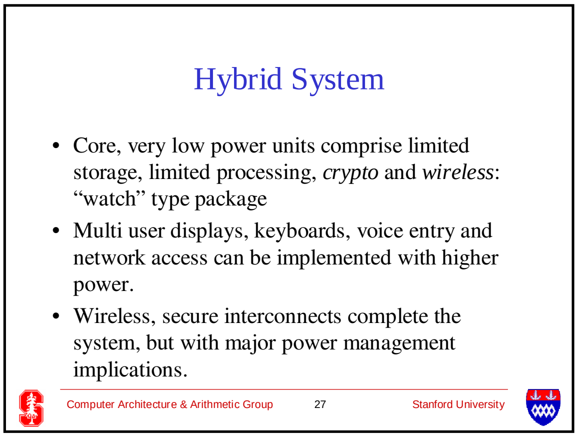
NO – small, embedded die which are package limited. (10-100Mtx/die)



Area

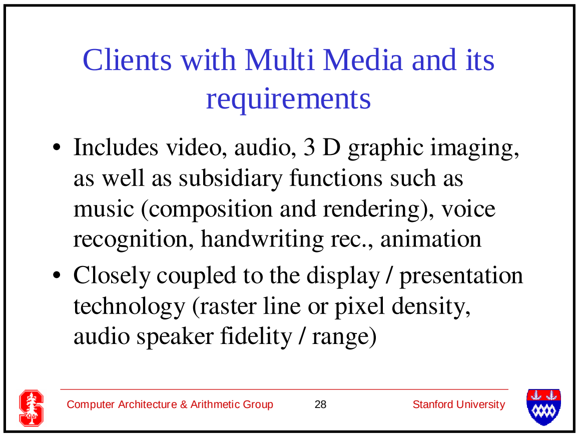
Indeed, looking ahead how to use 1-5 billion transistors/cm^2 die?

- With performance limited by power, memory & program behavior servers use multiple processors per die, limited by memory.
- For client systems either “system on a chip” or hybrid system are attractive.
- Low cost embedded systems need imaginative packaging and enhanced function.



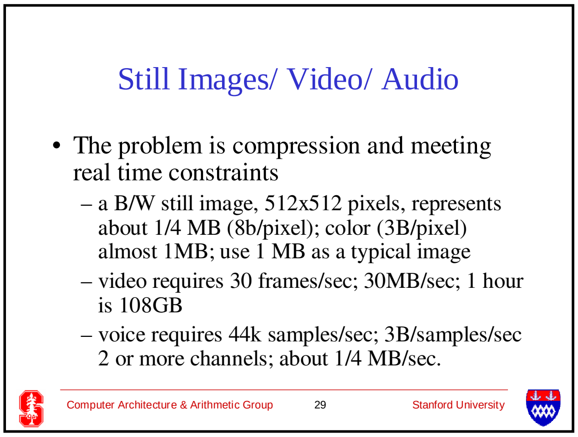
Hybrid System

- Core, very low power units comprise limited storage, limited processing, *crypto* and *wireless*: “watch” type package
- Multi user displays, keyboards, voice entry and network access can be implemented with higher power.
- Wireless, secure interconnects complete the system, but with major power management implications.



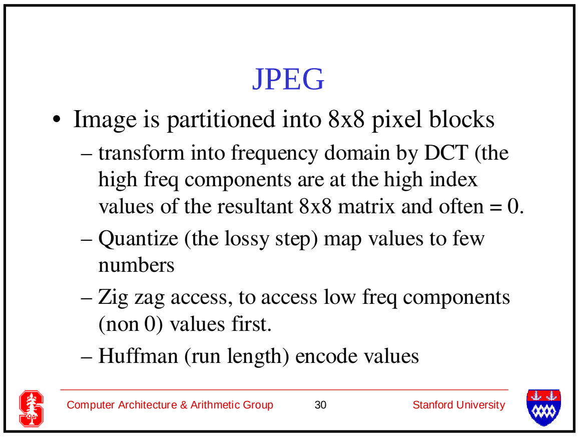
Clients with Multi Media and its requirements

- Includes video, audio, 3 D graphic imaging, as well as subsidiary functions such as music (composition and rendering), voice recognition, handwriting rec., animation
- Closely coupled to the display / presentation technology (raster line or pixel density, audio speaker fidelity / range)



Still Images/ Video/ Audio

- The problem is compression and meeting real time constraints
  - a B/W still image, 512x512 pixels, represents about 1/4 MB (8b/pixel); color (3B/pixel) almost 1MB; use 1 MB as a typical image
  - video requires 30 frames/sec; 30MB/sec; 1 hour is 108GB
  - voice requires 44k samples/sec; 3B/samples/sec 2 or more channels; about 1/4 MB/sec.



JPEG

- Image is partitioned into 8x8 pixel blocks
  - transform into frequency domain by DCT (the high freq components are at the high index values of the resultant 8x8 matrix and often = 0.
  - Quantize (the lossy step) map values to few numbers
  - Zig zag access, to access low freq components (non 0) values first.
  - Huffman (run length) encode values

Discrete cosine transform (DCT)

- Map X (spatial domain) to Y(freq. domain)
  - more compact representation, use 8x8 pixel blk
  - $y[u,v] = (4C(u)C(v)/n^2)\Sigma_j\Sigma_k x(u,v) \cos(2j+1)u\pi/2n \cos(2k+1)v\pi/2n$
  - $C(w) = 1$ for $w=1,2\ldots$ or $C(w) = 1/\text{sqrt}2$ for $w=0$
  - better than discrete Fourier transform but needs more computation
  - low precision

Video

- Popular standards:
  - H263 (video conferencing)
  - MPEG 1 (VHS quality)
  - MPEG 2 (Broadcast quality)
  - MPEG 4 (uses VOPs to achieve high quality with good compression)…. More complex, an emerging standard

Typical compression

- Image size, quality and delay are factors
- Lossless 3:1
- JPEG 25:1
- MPEG1 100:1 uses 352x288 CIF; 1-2 sec
- H 263 maybe 300:1; QCIF 176x144; 1/4sec
- MPEG2 4xCIF uses lower Q; longer delay

MPEG frames

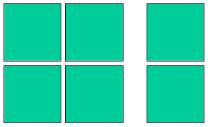
- Three types of frames
  - I intra-picture, like lossy JPEG
  - P predicted picture, motion prediction based on earlier I; motion vector plus error terms, as error terms are small quantizing gives good compression
  - B bidirectional pictures, motion prediction based on past and future I or P
  - result is GOP eg IPBBPBBPBBPBBPBBI

I frames

- In MPEG typically use 1 I per 15 frames
- In H263 maybe 1 I per 300 frames
- I frames take (maybe) 4x bits to represent than a P or B frame.

P frames

- Motion prediction is computationally intensive; based on macro blocks 2x2blocks
- 16x16 of luminance, 1 8x8 Cr,1 8x8 Cb, color is interleaved (called 4:2:0)



Motion estimation

- Computation intensive
- compute SAD for all neighboring macro block combinations (index by 1 pixel).

  $\Sigma [x_{i,j}-y_{i,k}]$ across all macro blocks
- find location that minimizes SAD

Instructions /pixel

- JPEG about 320 to compress; 280 to decode
- MPEG1 about 1100 to compress; about 80 to decode.
- Note problem in motion estimation; need 352x288 x1100x30 instr /sec = 3.3 GIPS for MPEG1 to compress.
- MPEG2 uses bigger frames; better motion estimation and color ….maybe….20GIPS

Video memory

- Even if we have enough arithmetic BW, memory (cache) access is a problem. A single CIF frame has 200 - 400 kB and won’t fit into a L2 caches less than (say) 1 or 2 MB. Worse is the behavior of the L1 D cache. There are NO hits after a line is used.
- Solution: pre fetch and stride prediction caches at L1.

3D graphics

- For client processors it’s the big gorilla of compute requirements.
- Image creation/ reconstruction/ 2 and 3D projection/ animation/ visualization
- must be interactive for image creation and smoothly(visually) manage motion and update dynamics

3D graphics

- Some basic requirements
  - smooth motion 15 frames/sec
  - frame size 1260 x 1024 down to 512 x 512
  - pixel 24b (3 colors, RGB) + 8b color control sometimes double buffered and z buffered (3D) up to 96b total.
  - frame size x pixel size = frame buffer
  - frame buffer also used to refresh video @ 30 or 60 Hz

Operations, simple 2D type

- Limited primitives: polygon, line, circle, ellipse
- ops: scale, translate, overlay and clip
  - typical op is S op D =D; op is replace, or, xor, and …applied to blocks of pixels
  - computations linear, $y = mx + b$, or quadratic $x^2 + y^2 = R^2$; mostly on index values….needs say 16b, sometimes 32b FP computation

Compute requirements

- Assume a very simple image.
  - 10k triangles with 100 pixels/ triangles
  - 1024x1024 RGB display updated 10 frames/sec
  - diffuse ambient illumination
  - shading, no shadows
  - no texture, fog, transparency..
  - no clipping from multiple objects

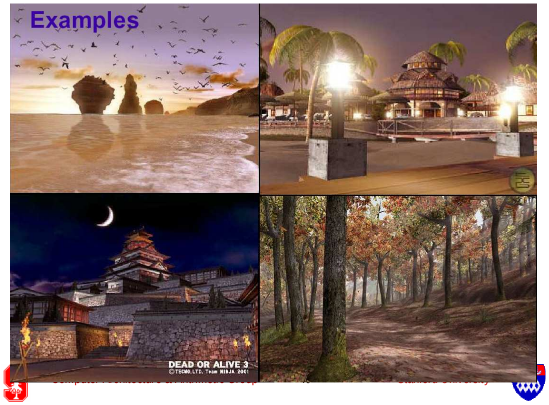
Compute requirements

- Transform/ rotate takes 25 multiplies& 18 adds/ vertex … 10k x 3 = 30k vertices
- computing viewer’s projection of the object surface (which triangles are in view) takes 18 multiplies & 14 adds / vertex
- simple lighting 12 mpy & 5 adds/vertex
- clipping 3 divides/ vertix
- plus….

Compute requirements

- Net about 2M multiplies &1.4M adds/frame
- about 30 Mflops depending on image clipping adds 1M divides/sec
- plus about 50 Maps to the frame buffer.
- Refined images (10 pixels/polygon) plus Phong illumination and Gourand shading plus shadows, texture, etc…. 100x to 400x
- WHEW

From Montrym (Nvidia) Hot Chips ‘02



Arithmetic for multi media

- Must support very high bandwidth arithmetic
  - pipelined integer and fp
  - optimized for 8, 16 and 32b operands
  - instructions set support for sub word concurrency
  - divide and sqrt can be important as well as trig functions.

Designing for multi media

- Must support very high bandwidth memory
  - structured memory access VRAM
  - VRAM on chip?
  - structured L1 D cache
  - large L2; maybe also structured or bypassed as with vector processor.

Beyond T x A x P …other dimensions

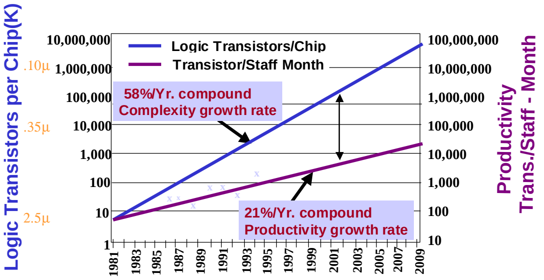
- Computational integrity and RAS
- Design time and FPL
- Configure ability / system connections and wireless technology.

Computational Integrity

Design for

- – Reliability
- – Testability
- – Serviceability
- – Process recoverability
- – Fail-safe computation

CAD productivity limitations favor FPL



Source: S. Malik, orig. SEMATECH

FPL (FPGA) and arithmetic

- Need robust compiled arithmetic units, scalable as to precision, area (CLBs), and time.
- While inefficient in latency can be quite useful for arithmetic bandwidth.

Wireless technology

- Essential to realizing efficient client and package limited systems
- Can be power consumptive, requires very efficient channel management.
  - – Focused, adaptive radiation patterns
  - – Digital management of RF
  - – Probably better suited to MHz rather than GHz RF… in conflict with antenna goals
  - – Many issues: security, multi signal formats, etc.

Regulatory Overview

- Free spectrum for WLAN has been allocated on a country/ regional basis
  - – 2.4-2.483 GHz (83 MHz total) has been available in most counties for many years
  - – Many countries have created much larger allocations at 5 GHz in the past few years:

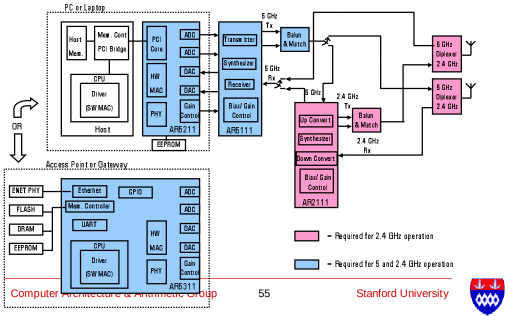
Region	5.15-5.25	5.25-5.35	5.47-5.725	5.725-5.825	Total (MHz)
USA	200mW	1W		4W	300
Europe	200mW	200mW	1W		455
Japan	200mW				100
Asia Pacific, including Korea, Hong Kong, China, Singapore, New Zealand, Australia	200mW (Singapore, Australia, NZ)	200mW (Singapore, Australia, NZ, Taiwan)		100mW to 1W (all, at various power levels)	300

- More spectrum = higher data rates and larger overall system capacity
  - – Radio is fundamentally a shared medium
  - – Current 2.4GHz systems are limited to only 3 independent 11 Mb/s channels
  - – Current 5 GHz systems have 13+ 54Mb/s channels

from McFarland Hot Chips '02

Atheros Chipset Overview

from McFarland Hot Chips ‘02



The “New” Server System

Server Processor

- Cost insensitive
- High power, fast cycle time
- Multiple (order of 10’s to 100) processors on single die, together with portion of memory space
- MIMD extensions via network
- Base processor using ILP with SIMD extensions
- Very high integrity

Client System

- Invisible core... integrated into
  - Appliances
  - Communications devices
  - Notebooks
  - “Watches”, wearable cores
- Always secure and available
- Very low power
- Integrated wireless/networked

The “New” Client System

Client

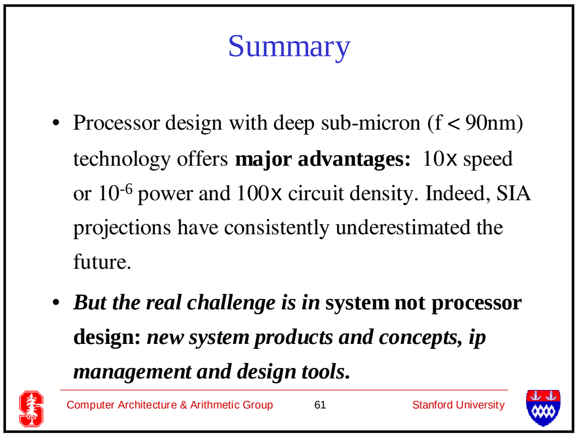
- “System on a chip,” integrated, hybrid or embedded
- Many frequency- power design points from 100 µw with cycle time 2 ns to 1 µw with cycle of 200 ns
- Small die ⟹ low cost
- Several base processors
  - Core and signal processors
- High integrity
- Signal processors VLIW/SIMD

System Design: The CAD Challenge

- IP’s – not IC’s. Amalgamate existing designs from multiple sources
  - Validate, test, scale, implement
- Buses and Wires
  - Placement and coupling, wire length, delay modules… some role for optics?
- Effective floorplanning

System Design: Other Challenges

- Really good power management
- Low power wireless interconnections
- Integrity – Security
- Scene/ pattern recognition
- Software efficiency: Op Systems, compilers
- Software productivity: rules based, analysis, heuristic, etc.



Summary

- Processor design with deep sub-micron (f < 90nm) technology offers **major advantages:** 10x speed or 10^{-6} power and 100x circuit density. Indeed, SIA projections have consistently underestimated the future.
- ***But the real challenge is in* system not processor design: *new system products and concepts, ip management and design tools.***