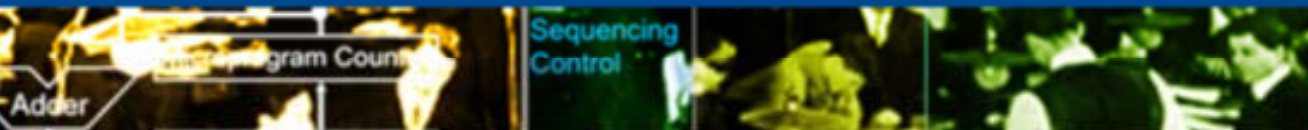


Reconfigurable Architectures for Desktop Supercomputing

by Daniel Alex Finkelstein
Quinnipiac University

April 19, 2005



Who am I?

Daniel Alex Finkelstein

PhD student

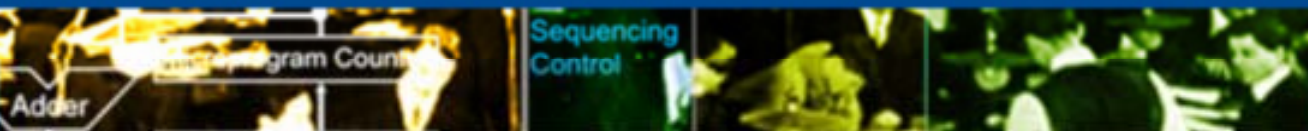
Research area: Computer Architecture

<http://cis.poly.edu/~dfinke01>

[Polytechnic University](#)

Brooklyn, NY

Adviser: [Prof. Haldun Hadimioglu](#)



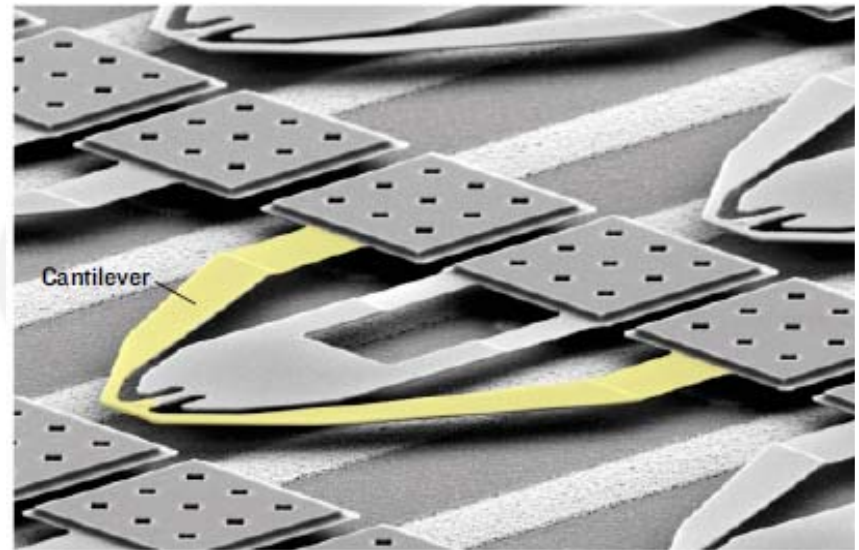
Problems in Architecture

1. The “Memory Wall”
2. Parallelism
 1. Multiple cores
3. Reliance upon Compilers & Operating Systems
4. Efficient ISA Support (RISC vs. CISC vs. EPIC)
5. Power density



Data Scaling

- Disk densities will increase to 1 TB / in² in 5 years
- 10+ Gb/s network throughput
- Desktop computers expanding beyond 2 GB of secondary memory



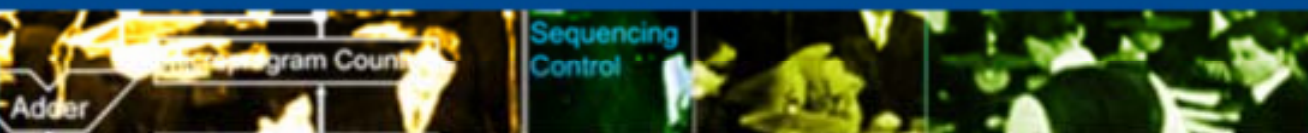
“Millipede”

H. Goldstein, “The Race to the Bottom,” in *IEEE Spectrum*, vol. 42, no. 3 (NA), March 2005, pp. 32-39.

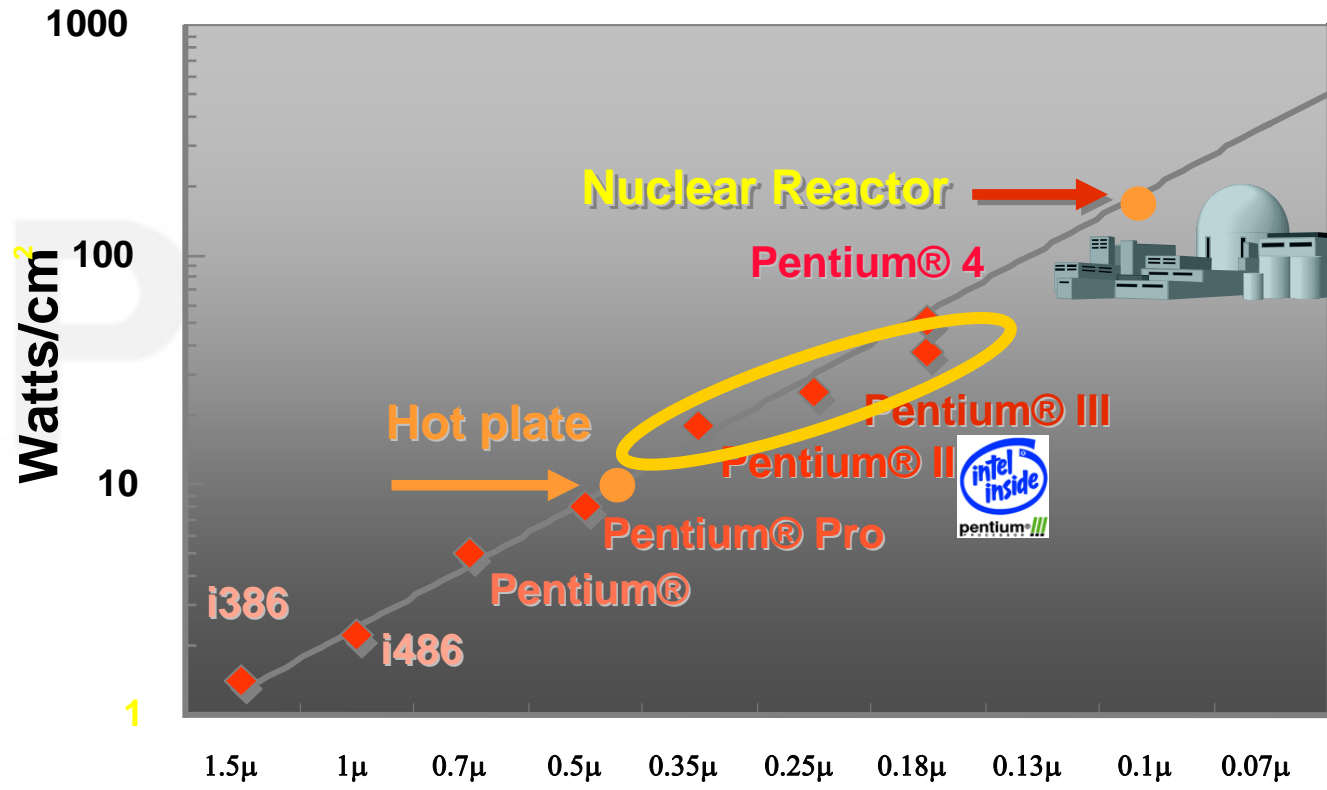


Hardware Predictions

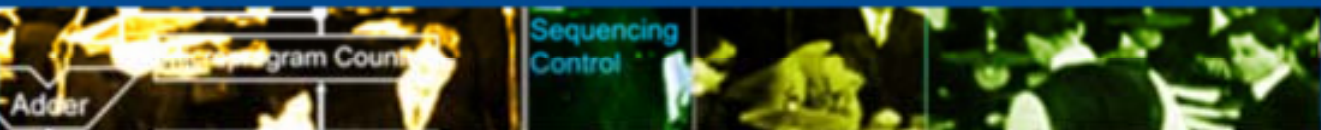
- Sematech's International Technology Roadmap for Semiconductors (2004) predicts that by 2018
 - There will be 14 billion transistors per chip
 - Chips will run at 53 GHz
 - 128 Gbit DRAM will be available



Power Density



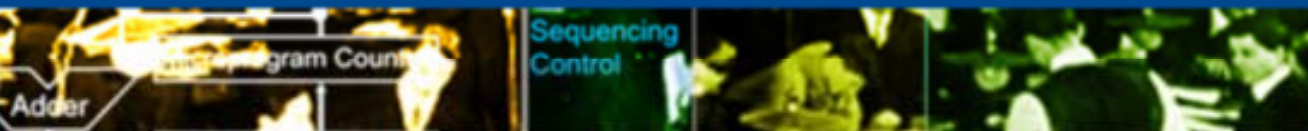
Dr. Avi Mendelson, Intel and Technion Institute,
<http://www.cs.technion.ac.il/~mendelson/Summary.ppt>



Goals

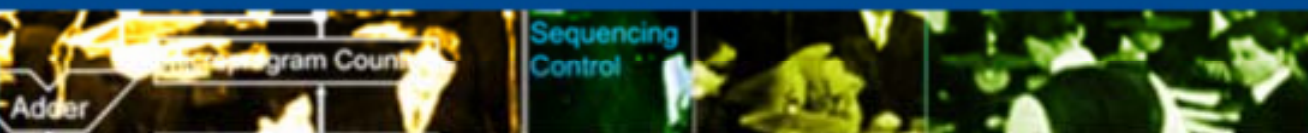
Supercomputer on a Desktop

- What applications matter?
- What trends will hardware follow?
 - “Today’s PC was yesterday’s supercomputer”



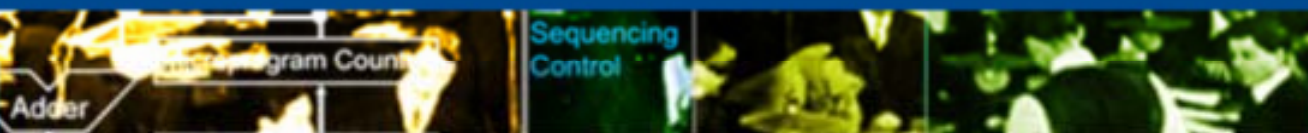
Candidate Applications for Supercomputing

- Computational Biology
 - DNA pattern queries
 - Protein Folding
- Weather Simulation
- Condensed-Matter Physics Simulation
- Image Manipulation
- Real-Time Signal Analysis (streams)



Candidate Applications for Supercomputing

- Gaming
- Multimedia Manipulation
 - iTunes: audio compression
 - DivX: video & audio compression



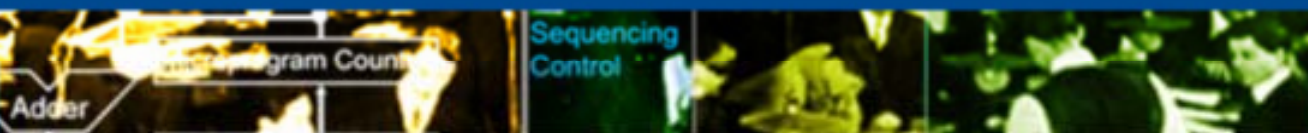
Memory Trends

- On-chip caches may be fastest (SRAM) but DRAMs are gaining ground.
- FCDRAM: Fast Cycle DRAM
- RLDRAM: Reduced Latency DRAM
- DDR2: Second-generation double data rate
- XDR: Octal data rate (Elpida)



Memory Products

Manufacturer	Type	Size (word format)	Latencies	Clock Cycle Time (min)	Random Access Time (max)
Elpida	DDR2	1 Gb (256Mx4)	4-4-4	3.75 ns	45 ns
	DDR	512 Mb (128Mx4)	2.5-3-3	7.5 ns	42 ns
	XDR	512 Mb (32Mx16)	2.0-2.5-3.33		28 ns
Toshiba	FCDRAM	512 Mb (4Mx8x16)		5 ns	22 ns



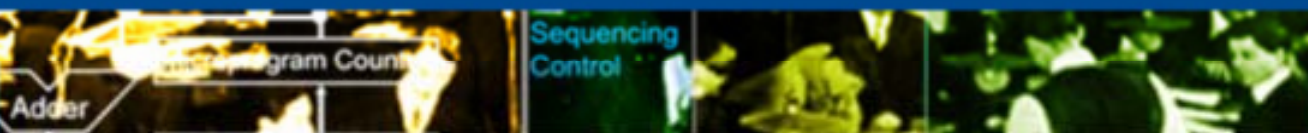
CPU Trends

- Processors range from the very tiny (PIC Microcontroller, ASIC) to the very large (Itanium2, Pentium IV, Athlon 64 FX, UltraSPARC IV)
- 1+ billion transistors are on the horizon.
- Coarser-grained parallelism:
 - ILP → SMT → CMP



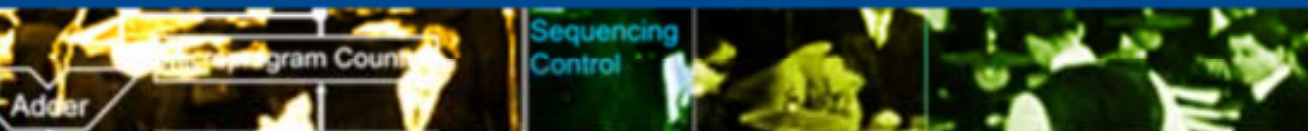
CPU Products

Manufacturer	Model	Bits	Cores	Threads	L2 Cache	Clock	Processor Bus
Intel	P4 Extreme Edition with HT	64	1	2	2 MB	3.73 GHz	1066 MHz FSB
	P4 with HT	64	1	2	2 MB	3.60 GHz	800 MHz FSB
	P4	32	1	1	1 MB	2.80 GHz	533 MHz FSB
	Pentium D	64	2	4	1 MB/core	3.2 GHz	800 MHz FSB
AMD	Athlon 64	64	1	1	1 MB	2.4 GHz	2 GHz HT
	Athlon 64 FX	64	1	1	1 MB	2.6 GHz	2 GHz HT
	Opteron	64	1	1	1 MB	2.6 GHz	1 GHz HT
	Opteron Dual	64	2	2	1 MB/core	2.6 GHz	1 GHz HT
Sun	UltraSPARC IV	64	2	2	16 MB/core (off chip)	1.2 GHz	150 MHz Fireplane



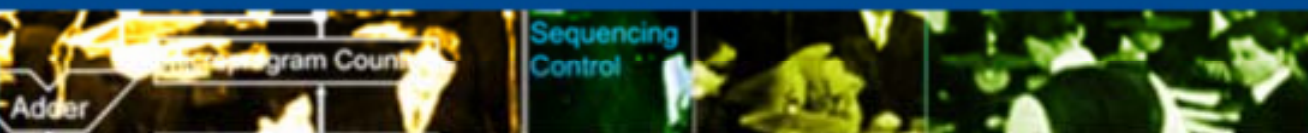
Trend CPUs

- “Niagara” from Sun
 - Will be the UltraSPARC VI
 - 8 cores handle up to 32 simultaneous threads (tiles)
- RAW from MIT (tiles)
- VIRAM from Berkeley (vector)
- Imagine from Stanford (streams)
- Cell from IBM, Sony, Toshiba (SMT)



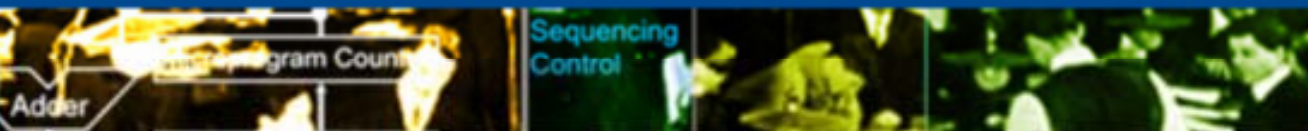
Computer Configurations

Manufacturer / Model	Processor	Clock Rate	# of Processors	Default Memory	Default Disk
Dell	Intel Xeon	3.60 GHz	2	4 GB DDR2 SDRAM	74 GB
Apple	PowerPC G5	2.5 GHz	2	512 MB DDR SDRAM	160 GB
HP Proliant	Intel Xeon MP	3.0 GHz	4	4 GB DDR SDRAM	36.4 GB
Cray XT3	AMD Opteron	2.4 GHz	30508	239 TB DDR SDRAM	n/a



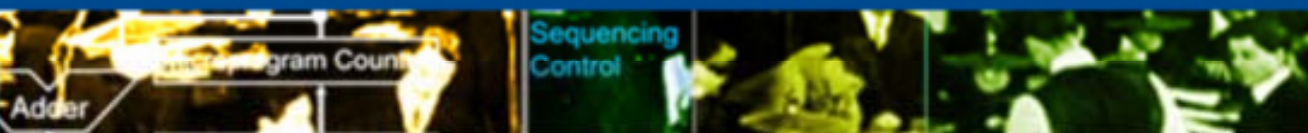
FPGAs

- Field Programmable Gate Arrays
 - Are reconfigurable
 - Run at hardware speeds
 - Often contain embedded processors
 - FPGA-fabric based, such as MicroBlaze
 - RISC core, such as PowerPC 405
 - Are relatively inexpensive



FPGA Products

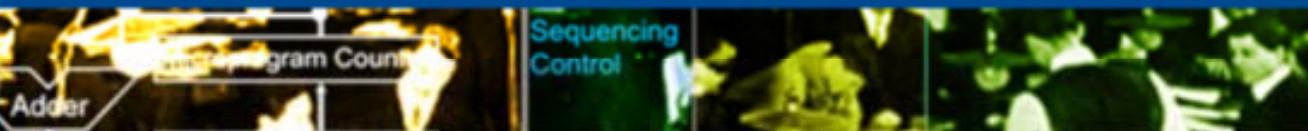
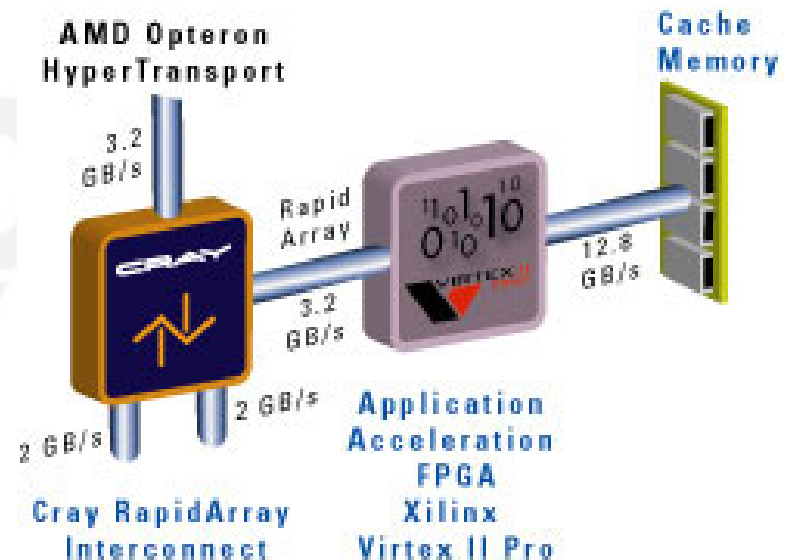
Manufacturer	Family	Device	Package	LCs	I/O	BRAM (Kbits)
Xilinx	Virtex 4	LX200	FF1513	220448	960	6048
	Virtex-II Pro	XC2VP30	FF896	30816	556	2448
	Spartan 3	50	VQ100	1728	63	72
Altera	Stratix-II	EP2S180	FBGA 1508	179400	1170	9163
	Cyclone-II	EP2C70	FBGA 896	68416	622	1125



Cray XD1

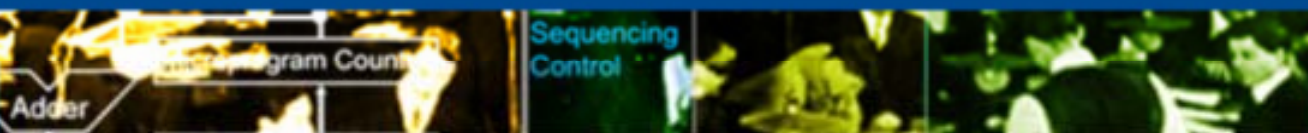


- Application acceleration through FPGAs (Xilinx Virtex-II Pro)
- Tightly-coupled host and FPGA
- Dedicated cache memory for FPGA (different from the Opteron's memory and cache)



Hybrid Execution

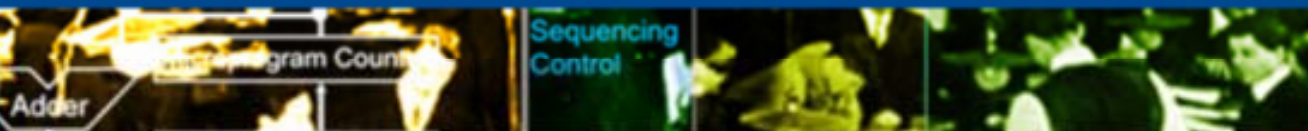
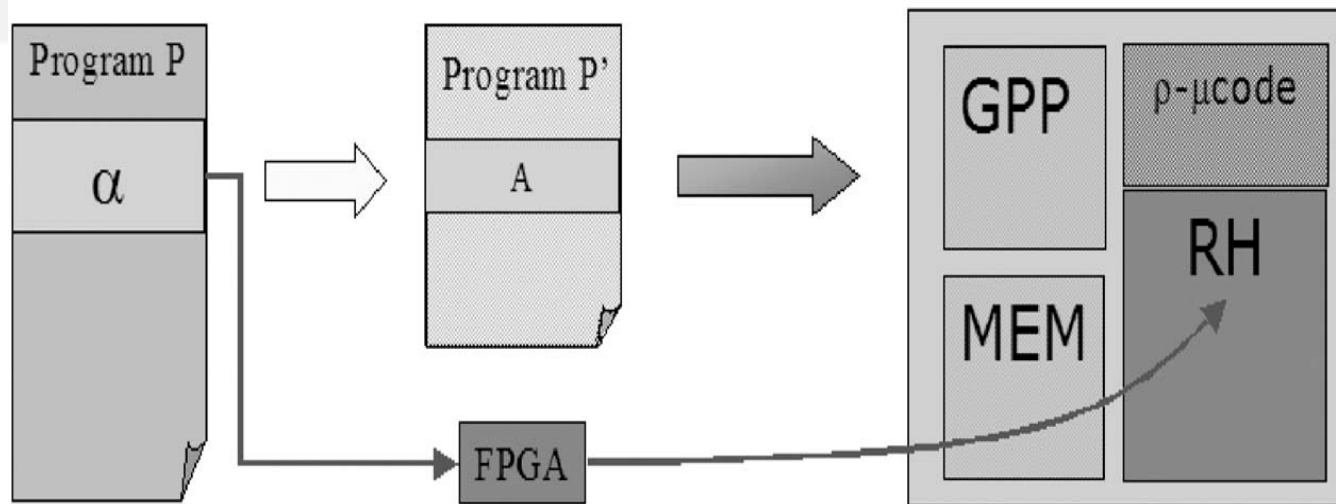
- Computationally expensive portions of code are sent to dedicated hardware for execution
- Requires new instructions and programming of the hardware (if reconfigurable) in advance



MOLEN

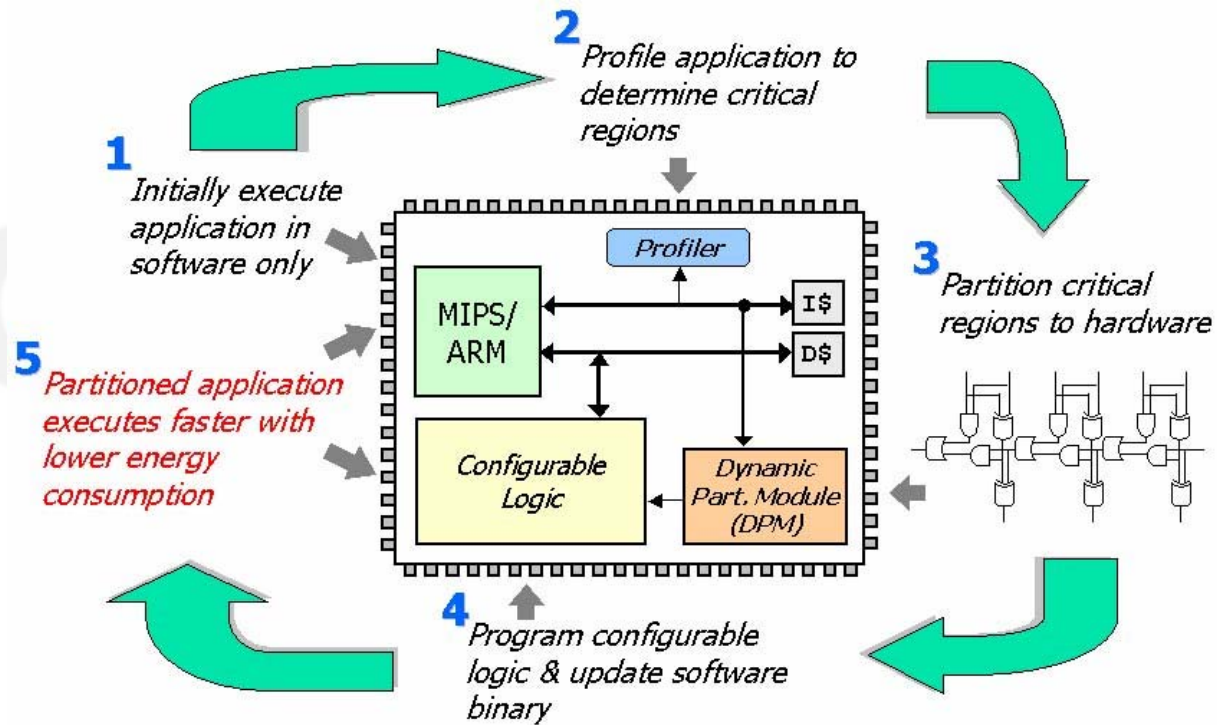
- Code α removed from program P and converted to hardware. Resulting program P' has stub A that refers to the reconfigurable hardware, which performs logic from α .

S. Vassiliadis, S. Wong, G. Gaydadjiev, K. Bertels, G. Kuzmanov, and E. M. Panainte. The MOLEN Polymorphic Processor. Transactions on Computers, 53(11):1363–1375, November 2004.

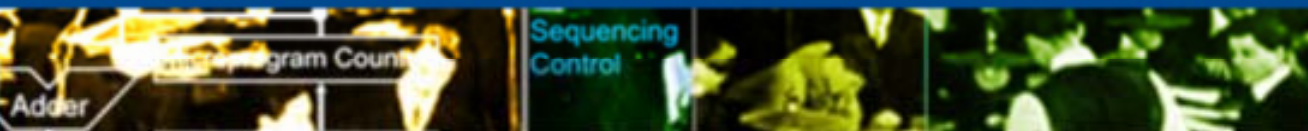


Warp

Continuous real-time profiling and reconfiguration

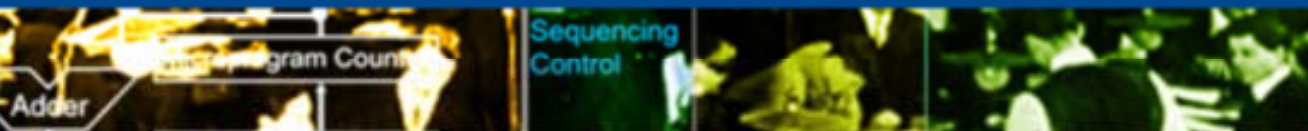


source: <http://www.cs.ucr.edu/~vahid/warp/>



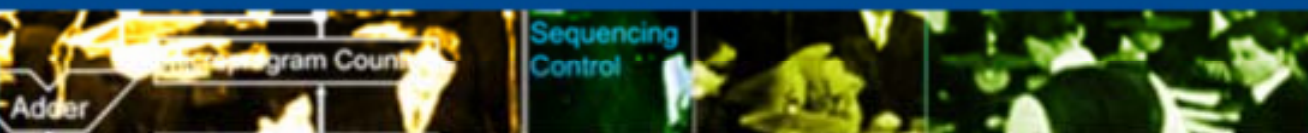
A New Approach

- Exploit regularity of scientific programs and parallelism of the algorithms
- Take complete control of secondary memory
- Present a CISC-like ISA to the world
- Adjust to user runtime requirements (speed, power, size)



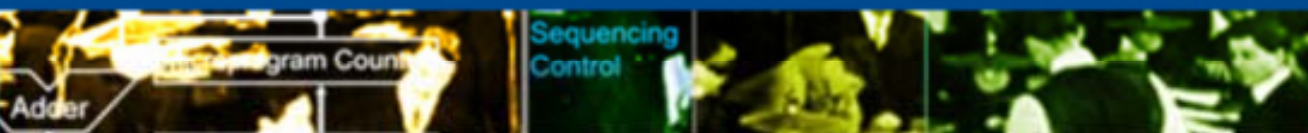
Memory Problems

- Speed of processors increases 60%/year while memory speed (latency) increases %10/year
- Interconnections (more pins, faster clocks) increase bandwidth, making the 'wall' more pronounced. Dual/multiple core even more so.
- Data structures should be intelligently placed in memory arrays for reduction of random access penalties.



Better Memory Controller

- An ‘intelligent’ processor to help the traditional processor (CPU) by
 - performing simple operations, like integer ops and compares
 - gathering data from secondary memory for upcoming operations, or reorganize memory contents to reduce random access penalties



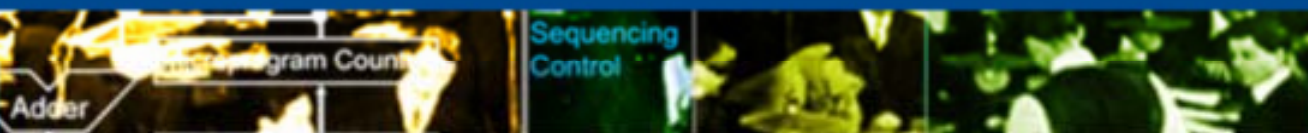
Plan Long Ahead

- Reconfigurable hardware can plan and schedule memory operations long in advance by having *a priori* access to the program code
- We reduce cache misses to the traditional processor
- Eliminate unnecessary speculative loads, or even speculative execution paths



In Conclusion...

- Computer architecture research is forging ahead as parallel processing becomes mainstream, multi-core chips come to market, very large memories are available to the masses, and extremely fast networks become ubiquitous.
- More exciting work ahead in architectural support for
 - software
 - networking
 - security
 - reduced power
 - memory
 - real-time
 - embedded systems
- And how do we encourage software engineers to think in terms of parallelism?



Questions?

Polytechnic
UNIVERSITY

dfinke01@photon.poly.edu

