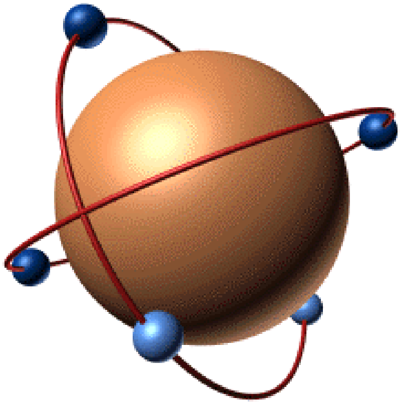
Heterogeneous Reconfigurable Systems

The Road to Low-power Systems-on-a-Chip

Jan M. Rabaey

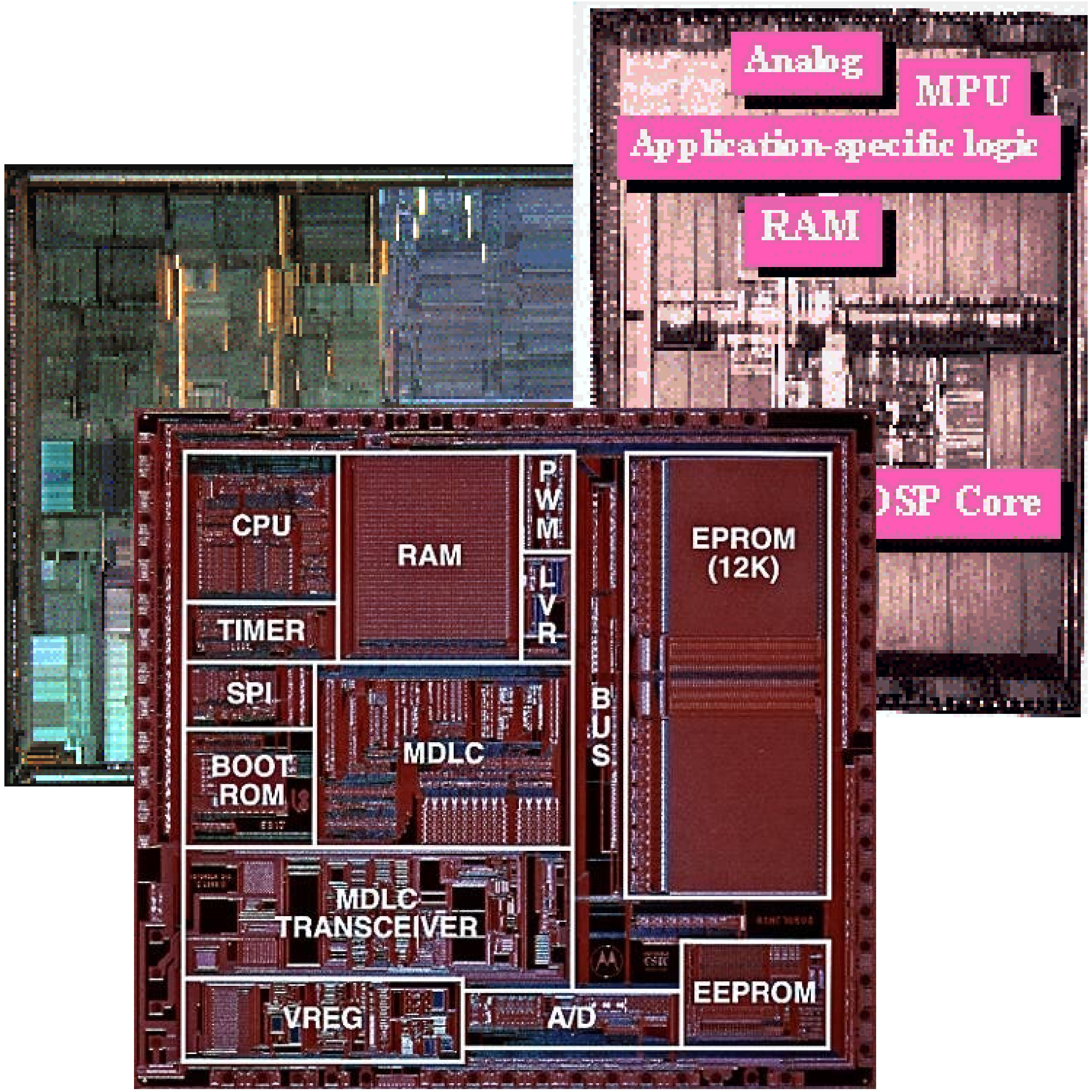
University of California @ Berkeley

October 31, 1997



http://infopad.eecs.berkeley.edu/research/reconfigurable

A new breed of designs



- Embedded applications where density, performance, and power are the real issues!
- DSP intensive
- Mix programmable and application-specific modules
- Mixed-mode

System-on-a-Chip

The System-on-a-Chip

"A system is a self-contained entity composed of a variety of components with heterogeneous properties communicating with each other using a variety of protocols" [VLSI93].

- Tracks the exponential growth in available transistors and interconnect
- A combination of pre-designed hardware modules and software physically realized as a single chip
- Subject to rigid constraints in Delay-Power-Area

Application Target

Small footprint, integrated embedded applications, that require high performance @ low energy. Programmability and adaptivity are essential

- Adaptive multimedia
- Multi-modal wireless radio’s
- Sensors or output devices integrated with sophisticated data formatting and processing

Opportunity: Most computational complexity and energy in a few kernels

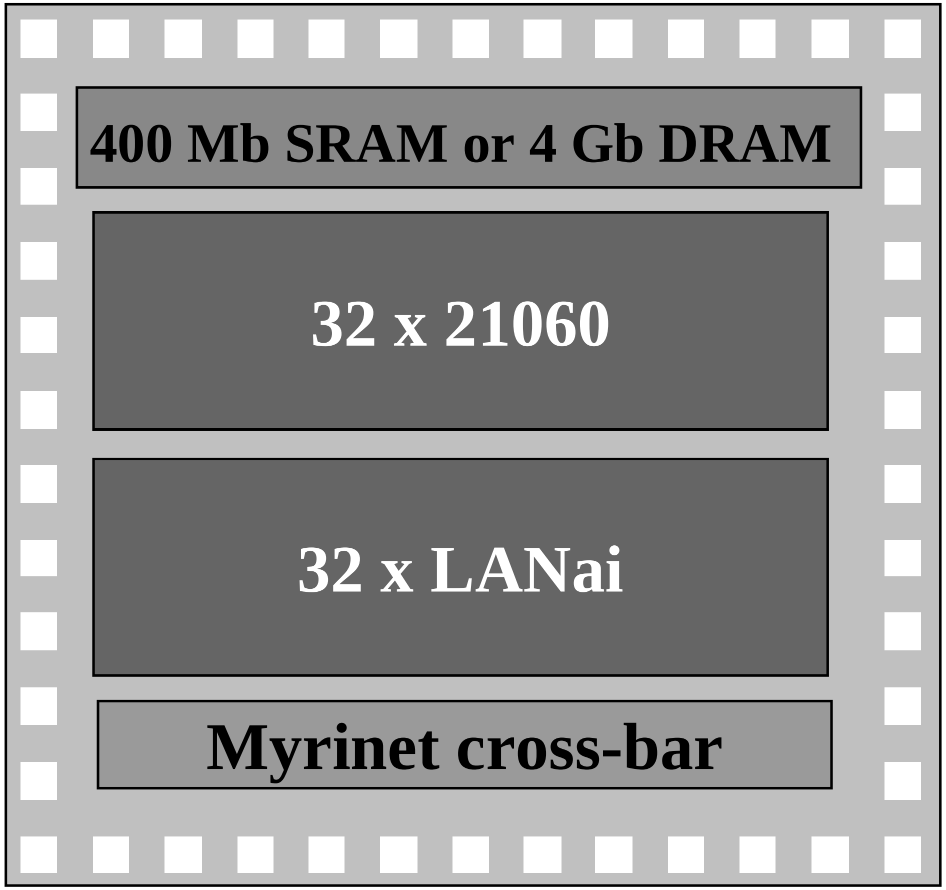
Perspective: Silicon in 2010

Die Area:	**2.5x2.5 cm**
Voltage:	**0.6 V**
Technology:	**0.07 μm**

	Density	Access Time
	(Gbits/cm2)	(ns)
DRAM	8.5	10
DRAM (Logic)	2.5	10
SRAM (Cache)	0.3	1.5

	Density	Max. Ave. Power	Clock Rate
	(Mgates/cm2)	(W/cm2)	(GHz)
Custom	25	54	3
Std. Cell	10	27	1.5
Gate Array	5	18	1
Single-Mask GA	2.5	12.5	0.7
FPGA	0.4	4.5	0.25

Example: SAR Image Formation



400 Mb SRAM or 4 Gb DRAM

32 x 21060

32 x LANai

Myrinet cross-bar

10MHz/ Watt!

Today

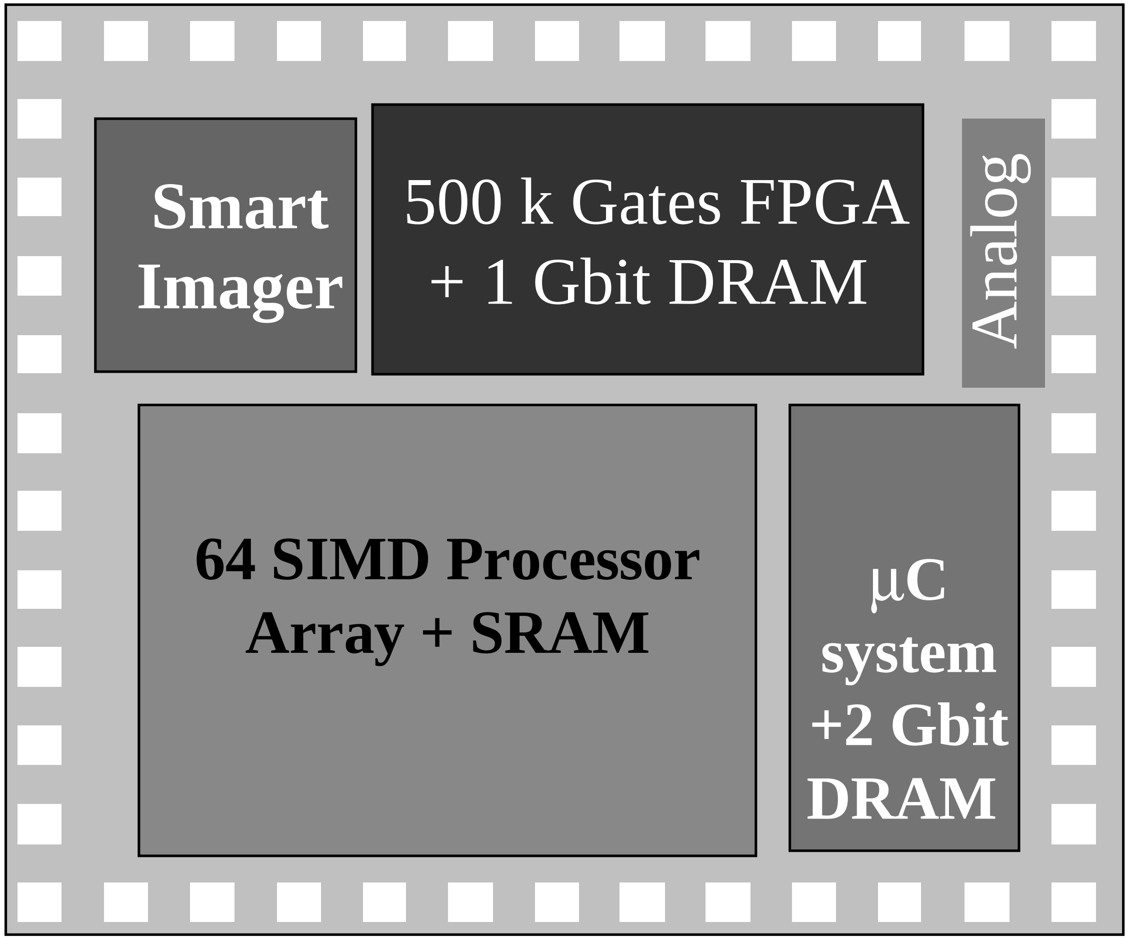
1 cubic foot
60 MHz - 1 kW
(0.6 MHz/Watt)

Integrated Sensor Systems

Multi-Spectral Imager: 1000x1000

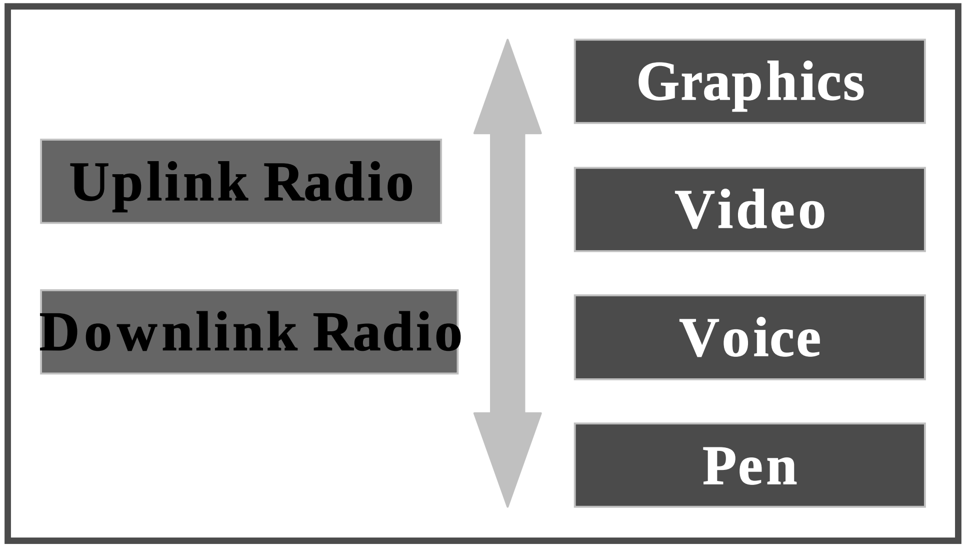
- **Visible**
- **IR**
- **Near-IR**

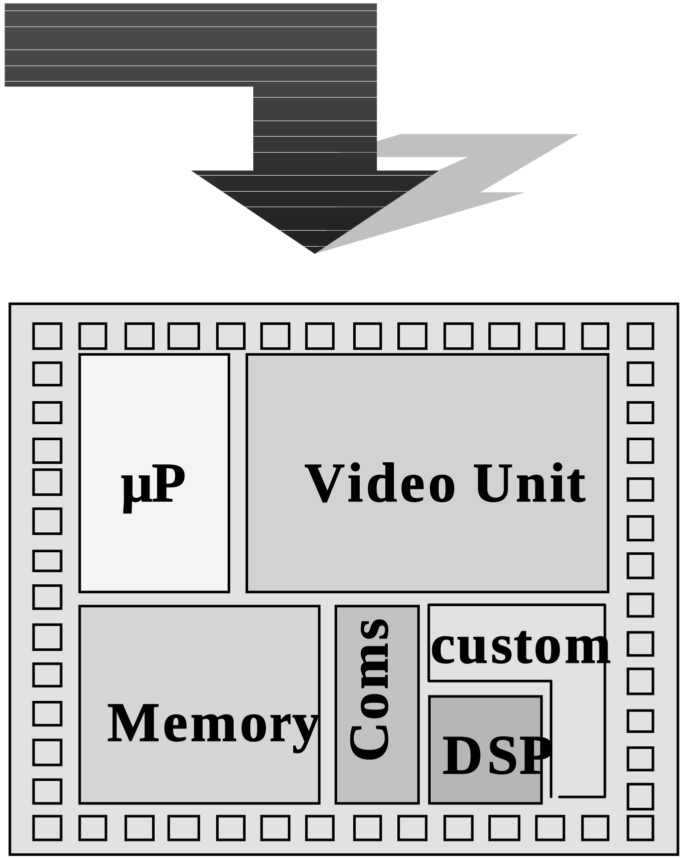
Image Conditioning: 100 GOPS


Smart Imager
500 k Gates FPGA + 1 Gbit DRAM
Analog
64 SIMD Processor Array + SRAM
μC system +2 Gbit DRAM

The SOC Opportunity

Mobile multimedia terminal

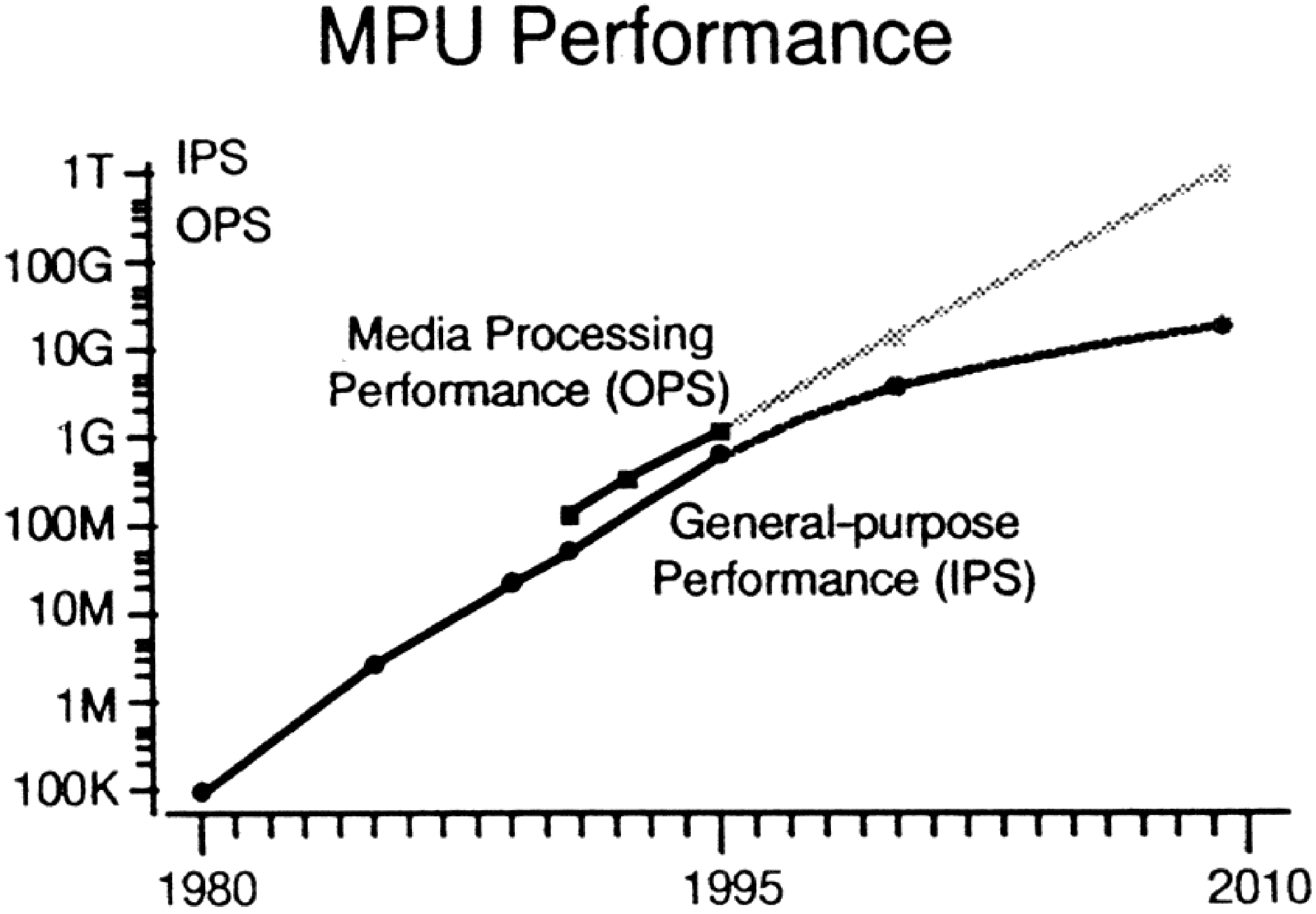




Opportunities for Optimization

- System Functionality
- Design Partitioning
- Architecture Selection

Motivating Heterogeneity



[Sasaki, ISSCC96]

Motivating Heterogeneity

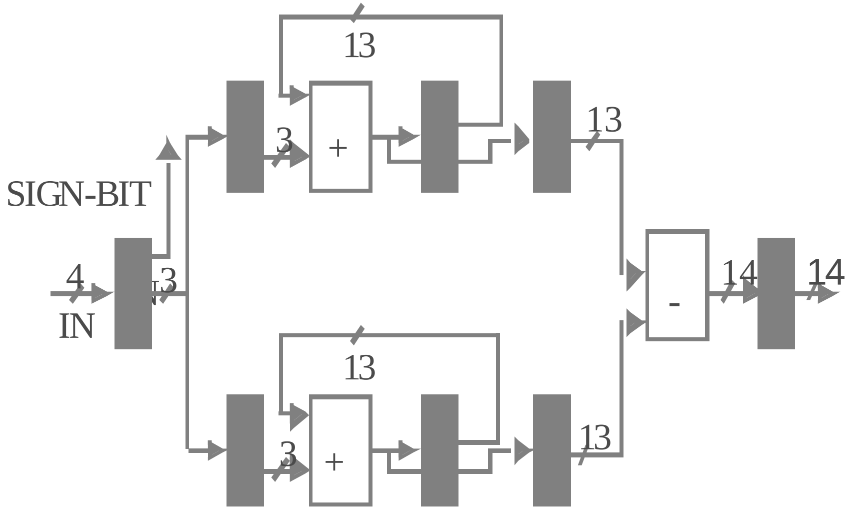
The Low-Energy Roadmap

- Voltage as a Design Variable
  - Match voltage and frequency to required performance
- Minimize waste (or reduce switching capacitance)
  - Match computation and architecture
  - Preserve locality inherent in algorithm
  - Exploit signal statistics
  - Energy (performance) on demand

- Easier accomplished in application-specific than programmable devices
- Requires new look at programmable architectures

Programmable versus Application-Specific

Example:
Correlator for CDMA Radio:



Energy/Flexibility Tradeoff’s

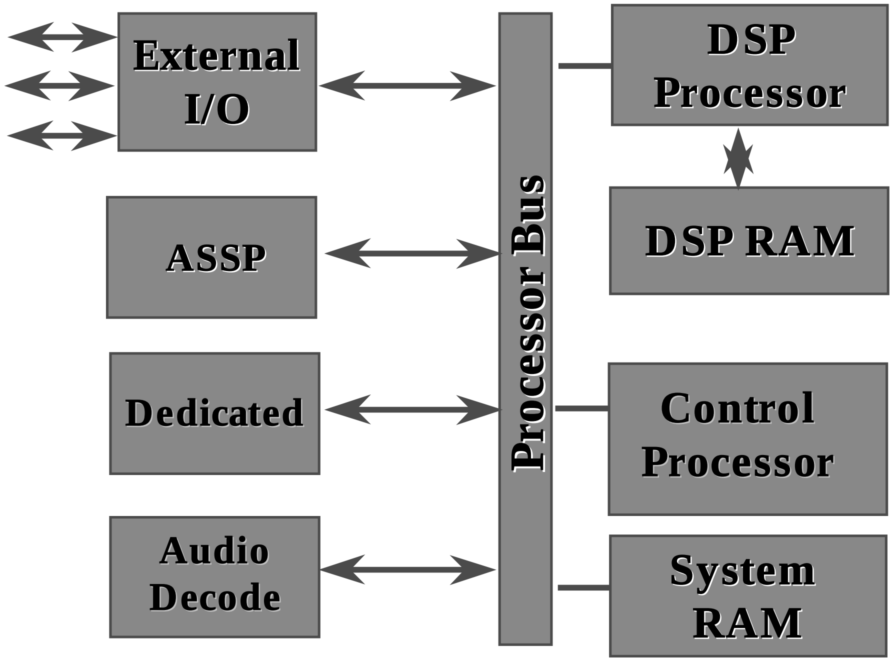
Arm 6 core (5V, 20 MHz):
2765 nJ 167697 fJsec

Xilinx 4003 (5V, 64 MHz)
394 nJ 394 fJsec

ASIC Datapath (1.5V, 64 MHz)
1.2 nJ 1.04 fJsec

* Energy/symbol
* Normalized Energy-Delay Product (5V)

A New Look at Architectures

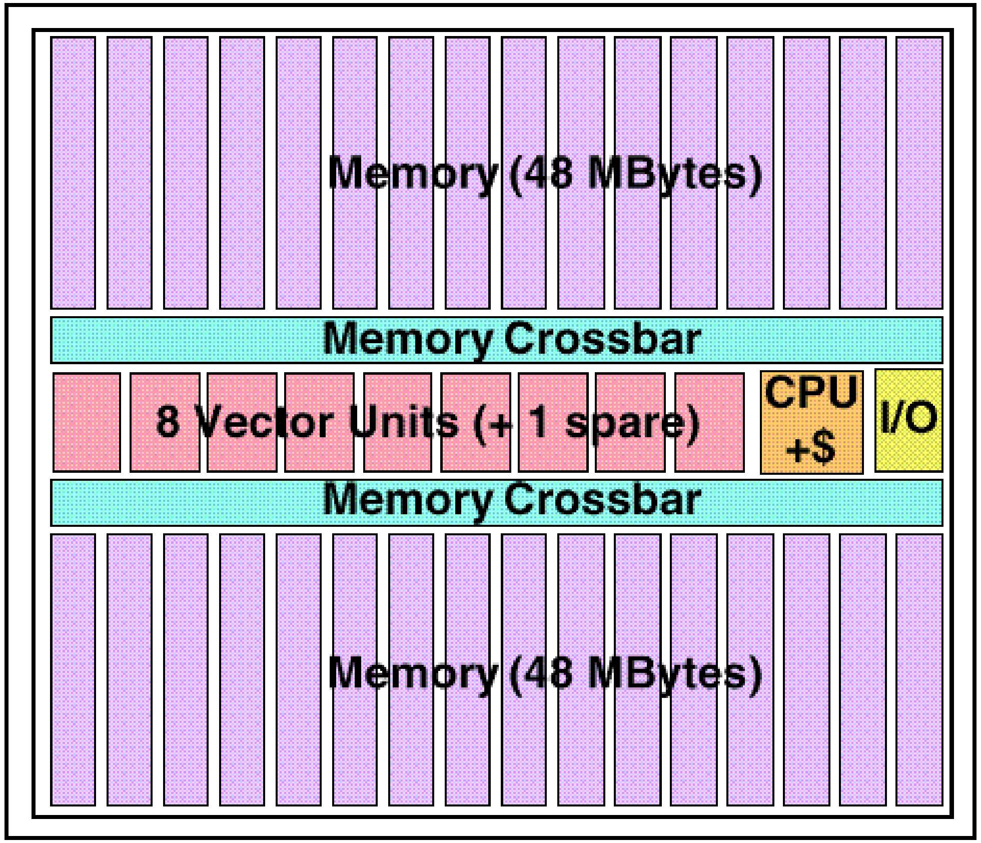


Application-Specific System-on-a-Chip

- Most common model at present
- Efficient
- but … ad hoc and $$$

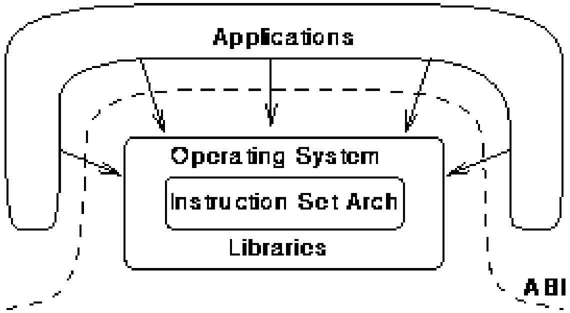
A New Look at Architectures

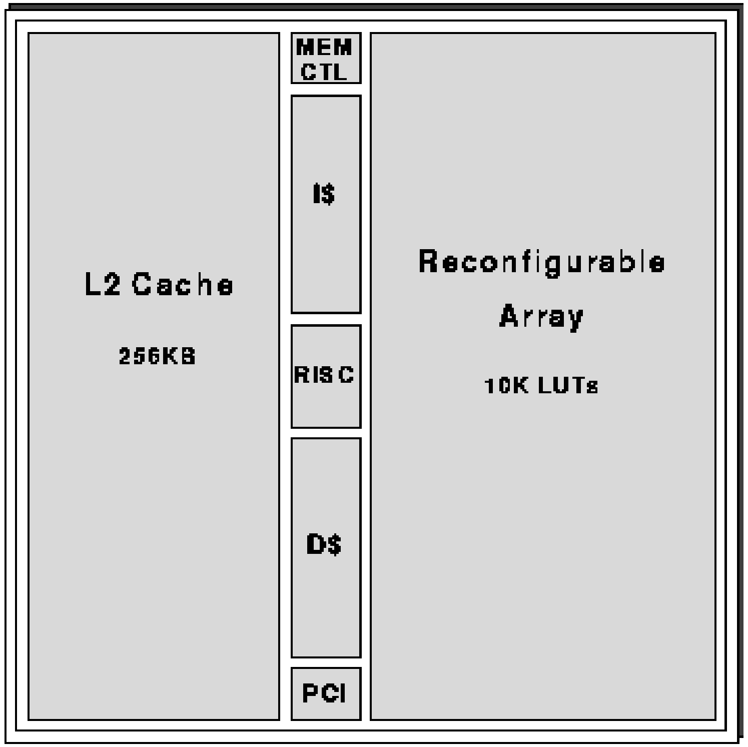
Example: V-IRAM-2 [Patterson97]: Combined RISC/ Vector Processor / DRAM



- 0.18 µm technology (2002)
  16 GFLOPS/128 GOPS
- Higher performance at lower clock due to parallelism
- Reduced overhead of instruction fetching (compared to VLIW and superscalar)
- Locality
- Optimized for application range (MMX)

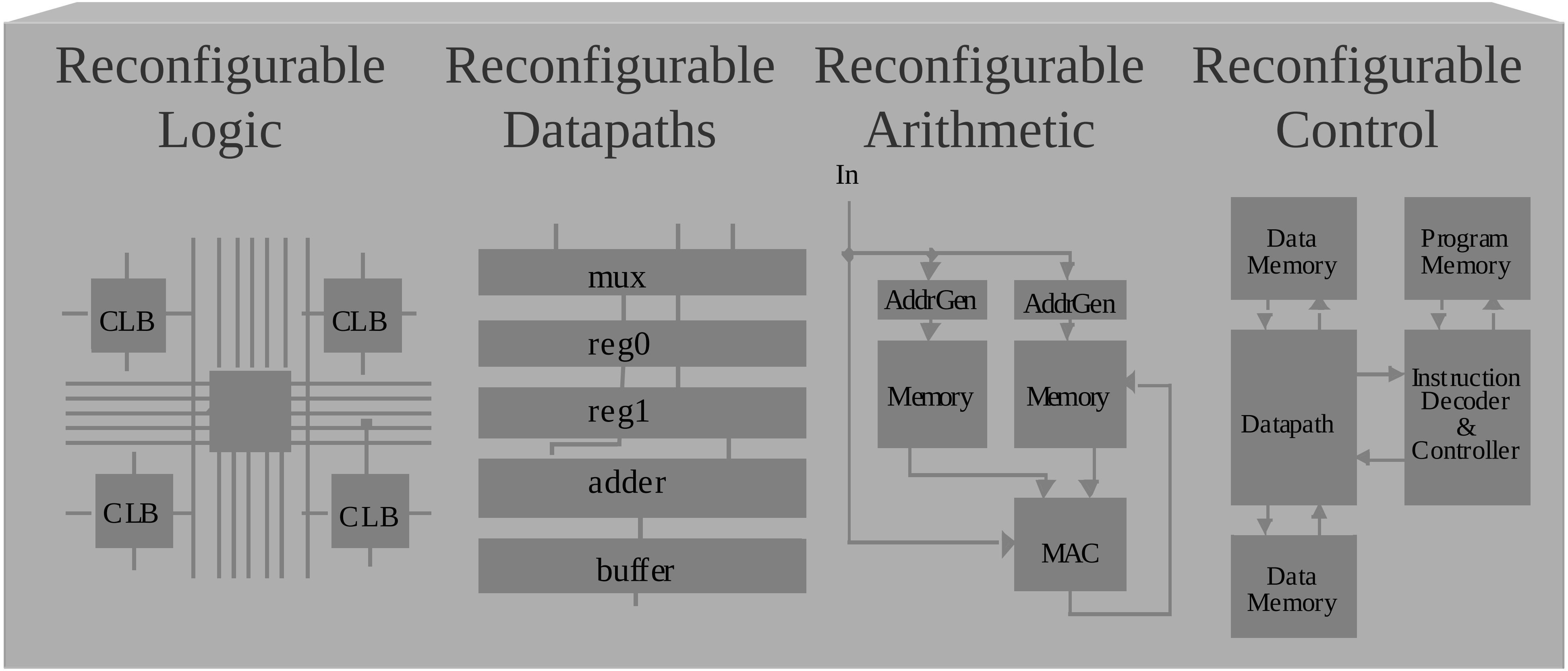
Merging RISC and FPGA





Examples: BRASS [UC Berkeley], Napa1000 [National]

A New Look at Architectures — Reconfiguration



Bit-Level Operations
e.g. encoding

Dedicated data paths
e.g. Filters, AGU

Arithmetic kernels
e.g. Convolution

RTOS
Process management

Challenges

- Understand and quantify the inherent advantages in terms of energy, performance and area of implementing an algorithm on a particular programmable fabric
- Develop good performance models to guide partitioning between heterogeneous programmable devices

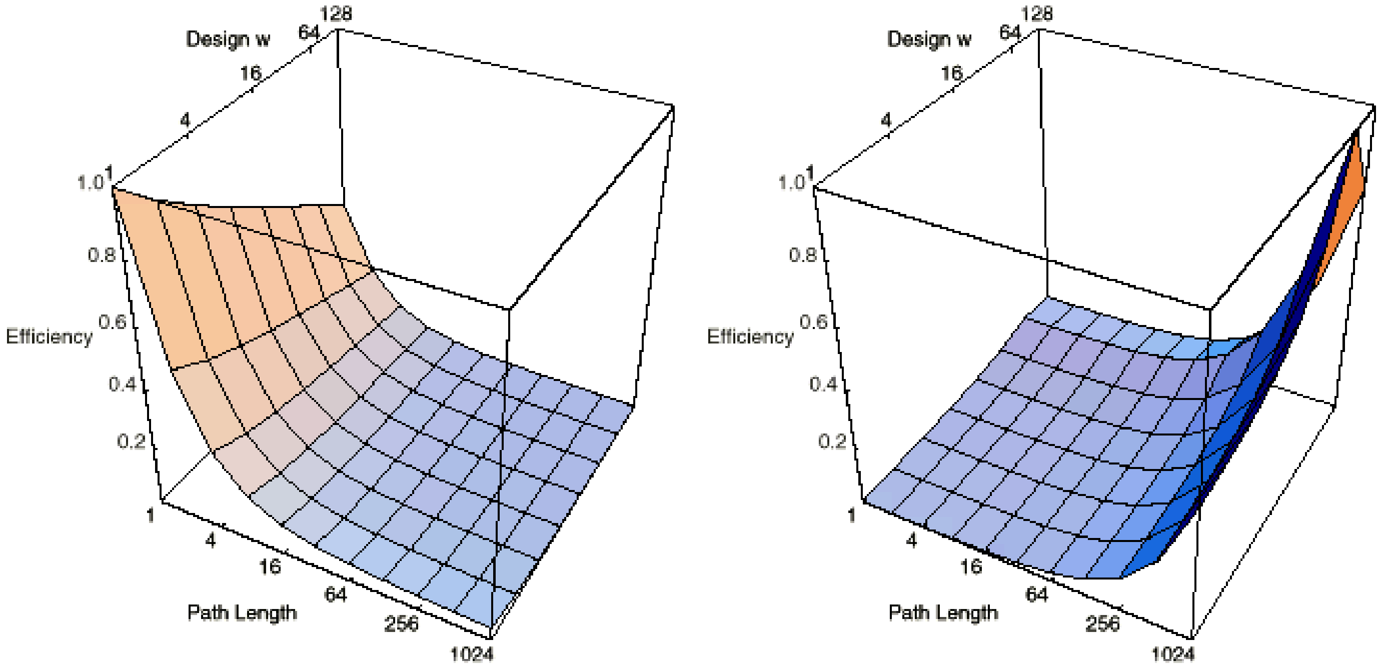
Roads to success:

  - » Benchmark analysis [Hennesy & Patterson]
  - » Parameter identification [Guerra95]

Architecture Parameters

- Programming Element
  - Granularity — size of PE in terms of word length, operators, data storage, contexts
  - Flexibility — range of operations that can be performed on PE
- Implementation Fabric
  - Homogeneity — variation of granularity and flexibility over PEs
  - Connectivity — degree of interconnectedness between PEs (includes locality and regularity)

Choice of granularity



FPGA

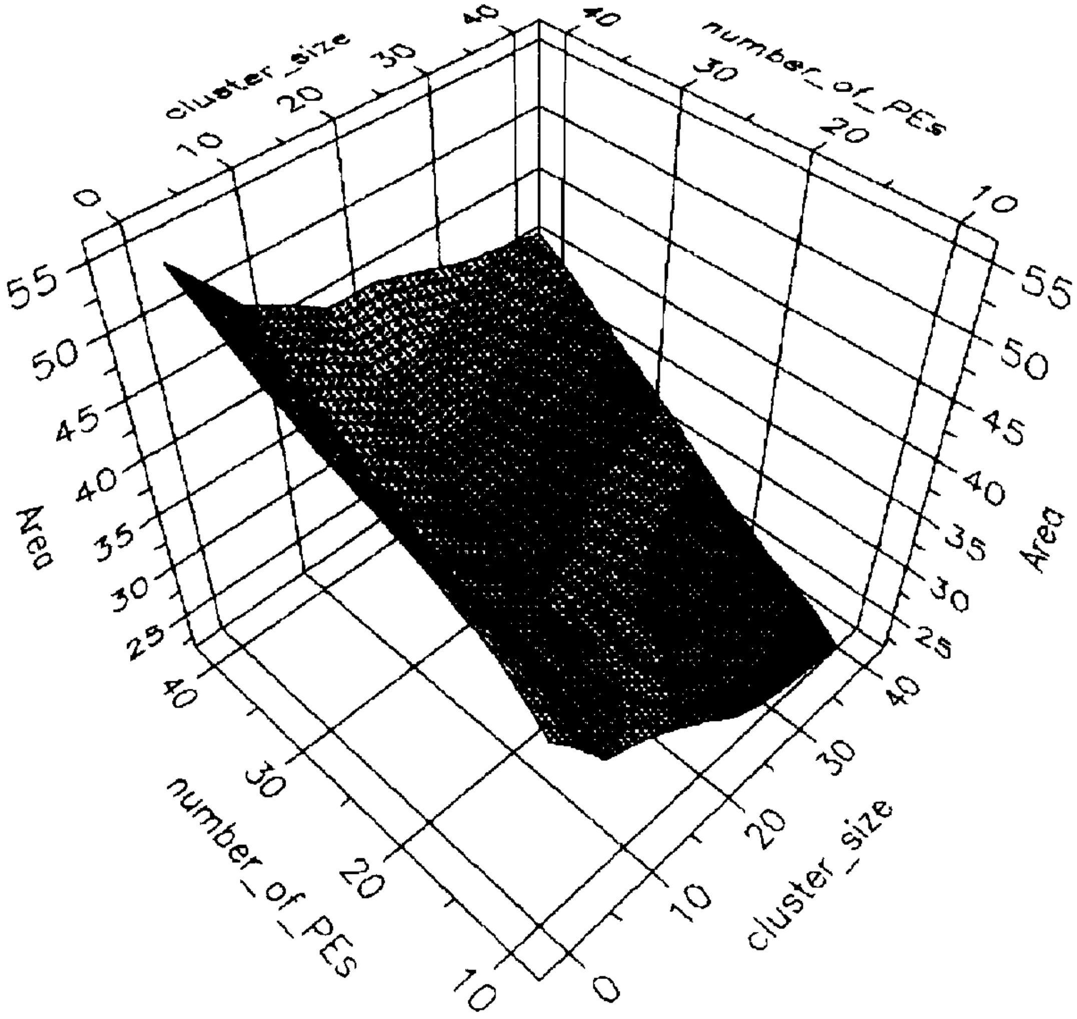
$c = d = 1, w = 1$

"Processor"

$c = d = 1024, w = 64$

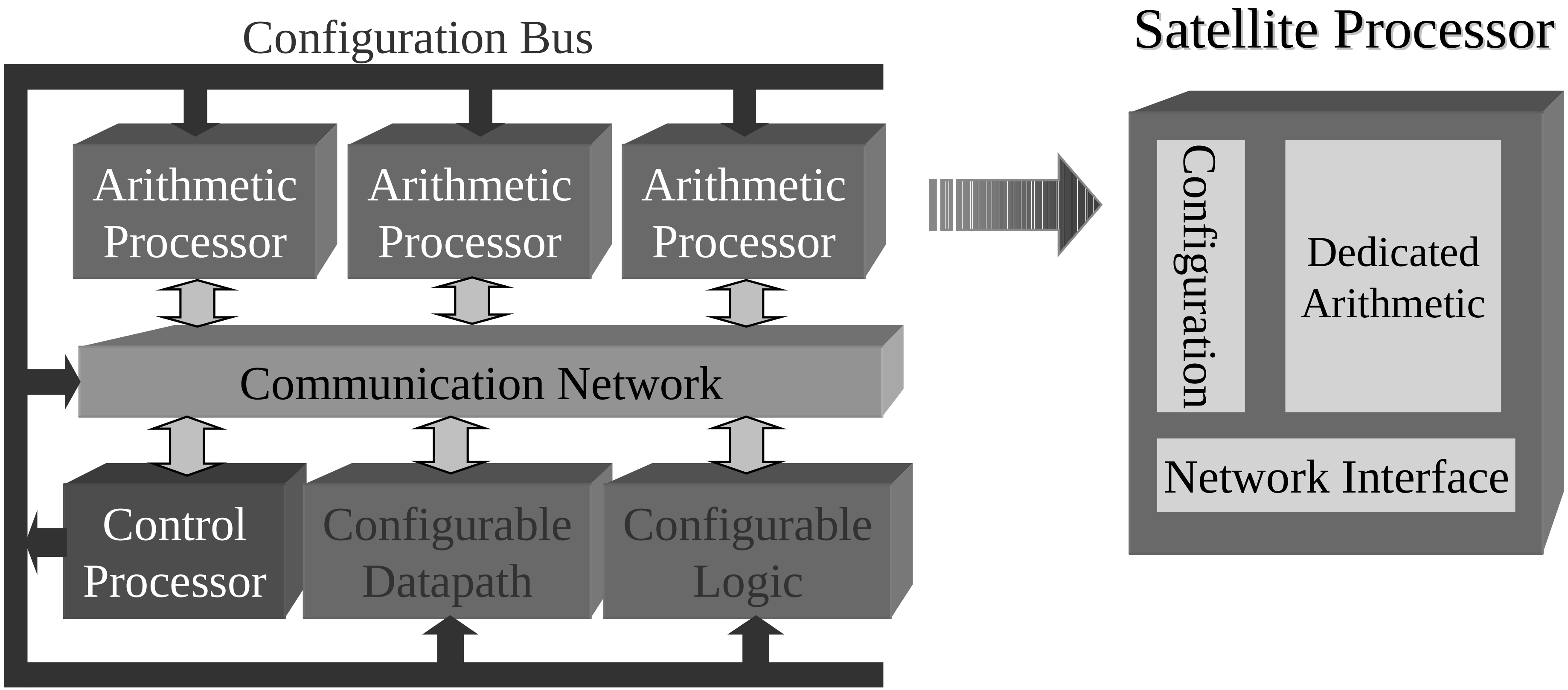
Efficiency comparison between processors and FPGAs (Dehon96)

Choice of granularity



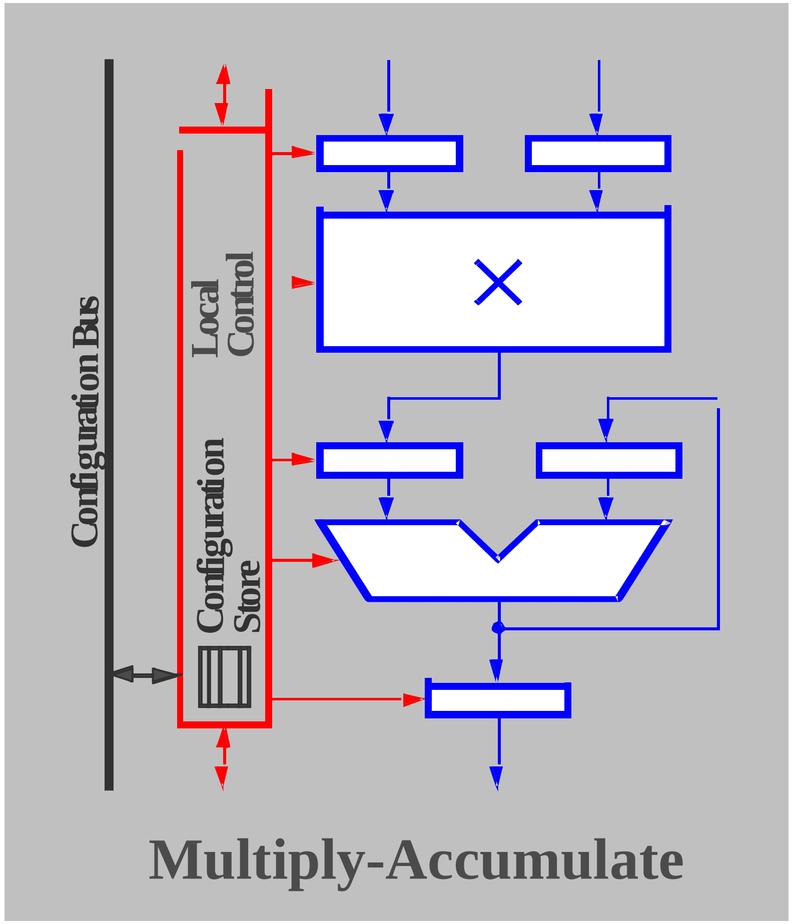
- Parametrical comparison over wide range of real-time video processing applications [Philips97]
- Uses reconfigurable array of weakly programmable processing elements
- Coarse-grain architectures are more efficient than fine-grained structures (for this class of applications)

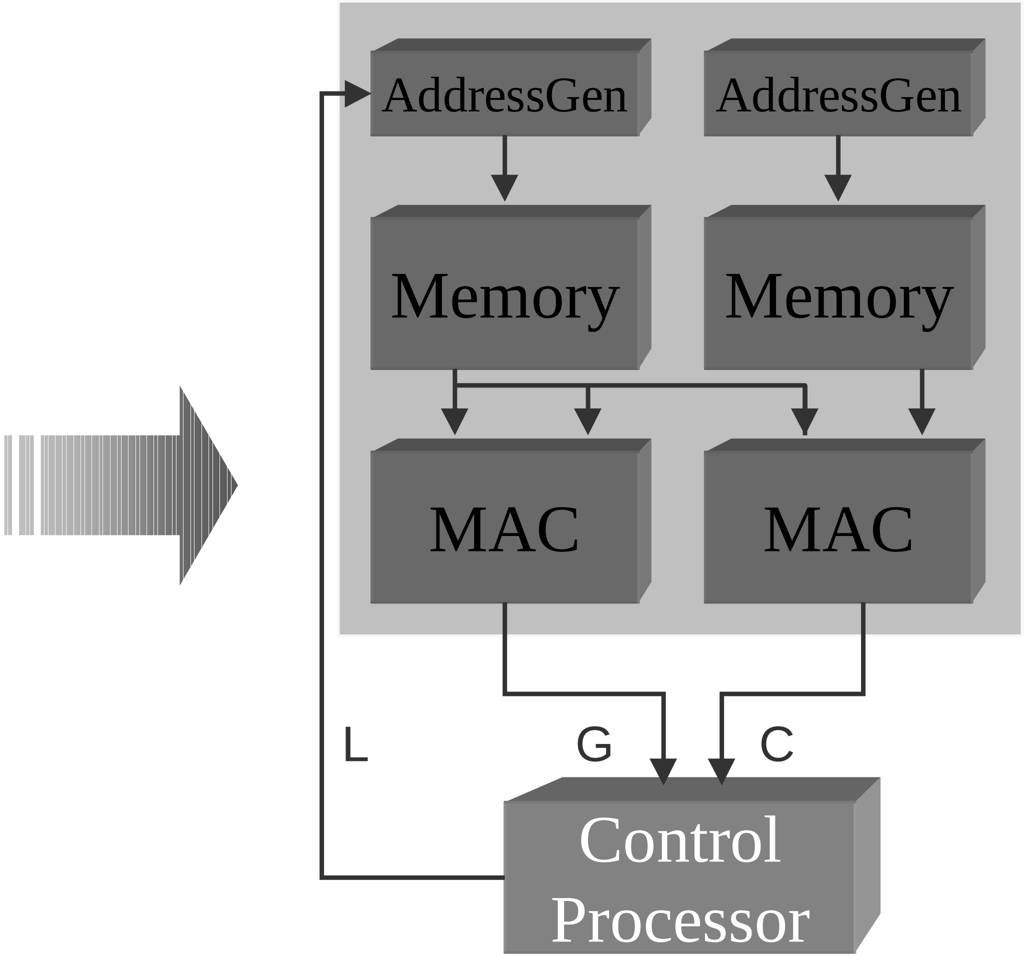
Multi-granularity Architecture — Berkeley Pleiades Project



- Computational kernels are “spawned” to satellite processors
- Control processor supports RTOS and reconfiguration
- Embodies all aspects of “Low Energy Roadmap”

Satellite Processors





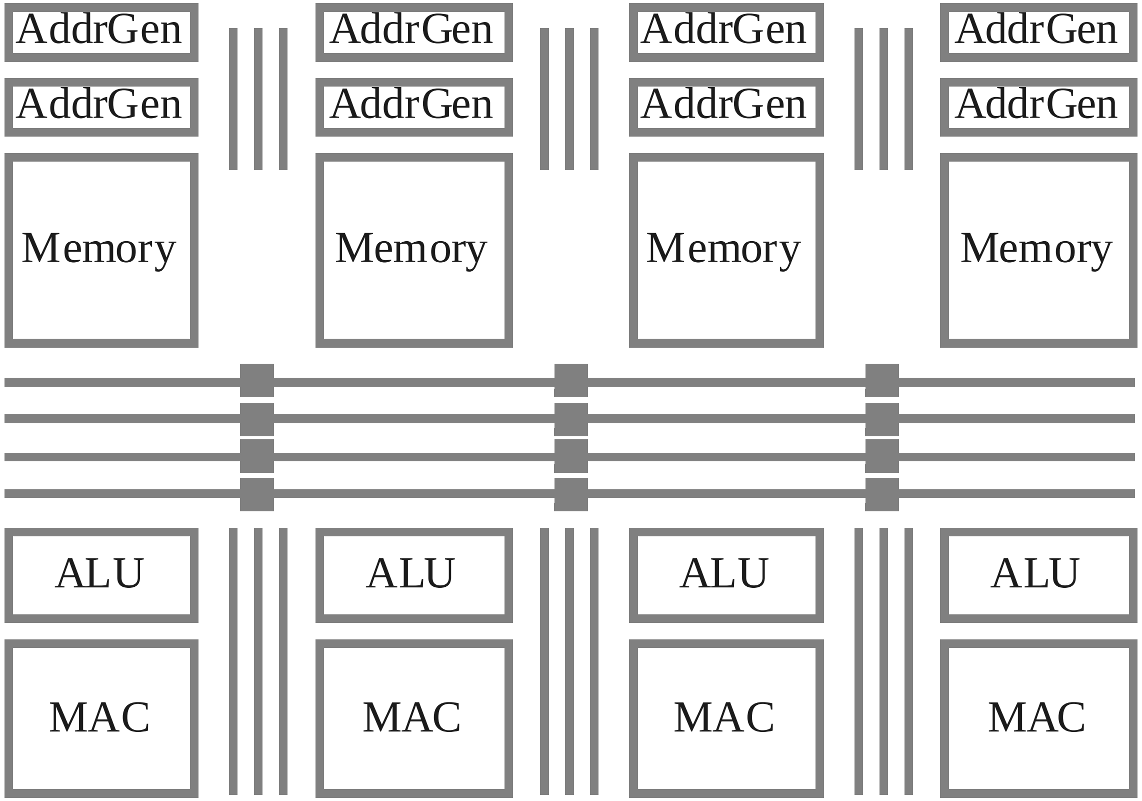
Communication Network

Dedicated links

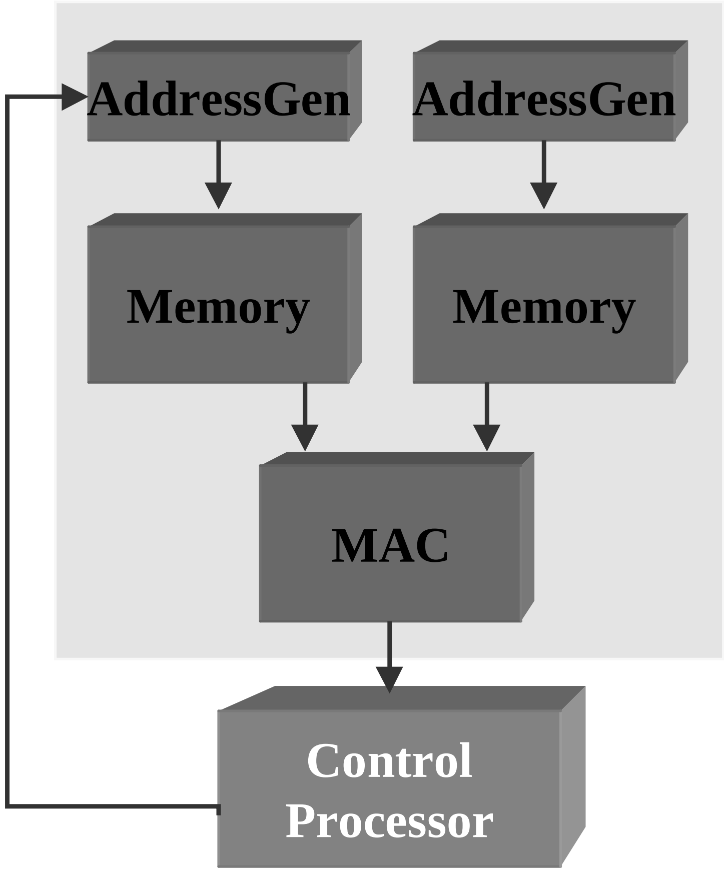
Reduced swing

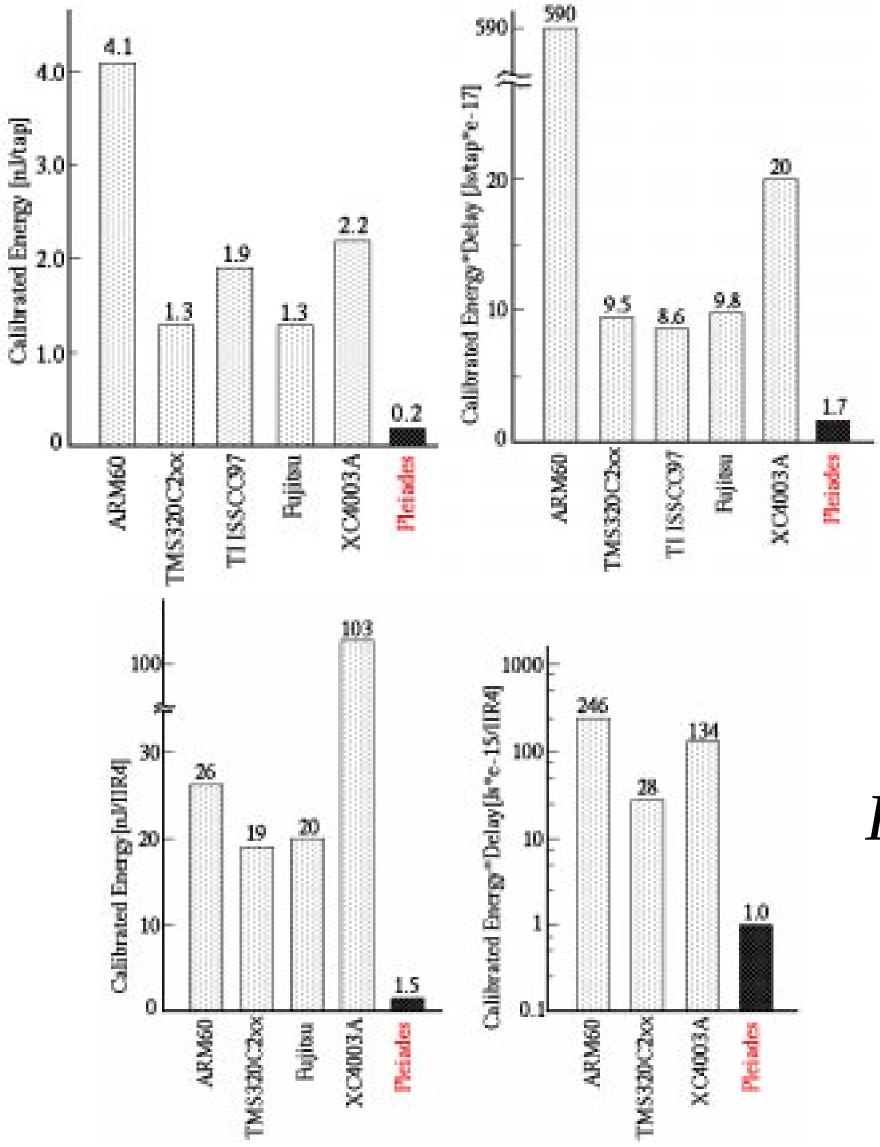
Local buses

Segmented buses



Architecture Comparison





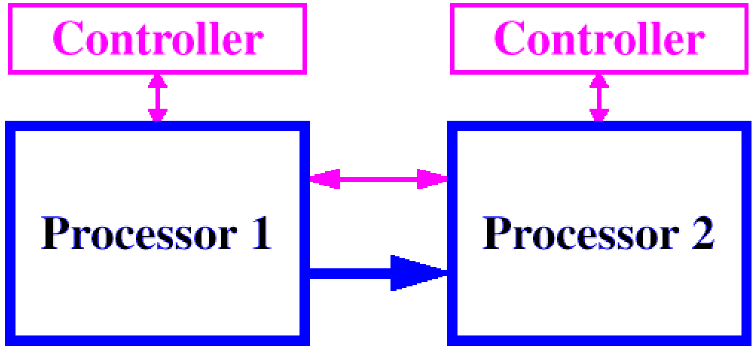
FIR

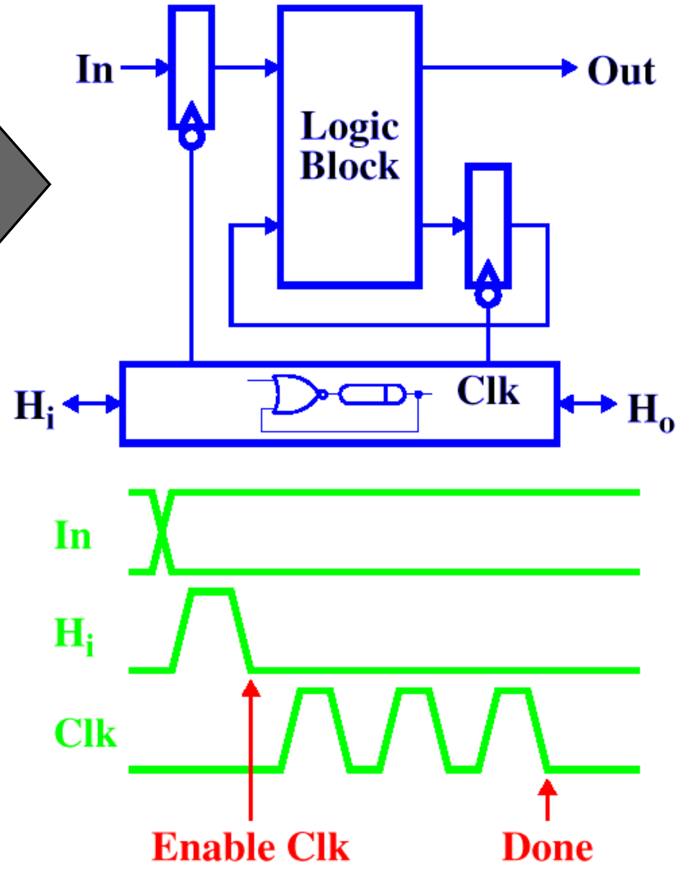
IIR

Aggressive Low-power Design

- Satellite processors optimized for specific task
  - » small control and data access overhead
  - » parallelism and pipelining easily supported
- Data-driven synchronization opens door for drastic reduction in clock power and enables globally asynchronous strategy
- Dynamic scaling and selection of supply voltage and execution rate using integrated dc-dc converter
- Low-swing configurable interconnect network
- Small grain FPGA cell reduces interconnect power

Synchronization

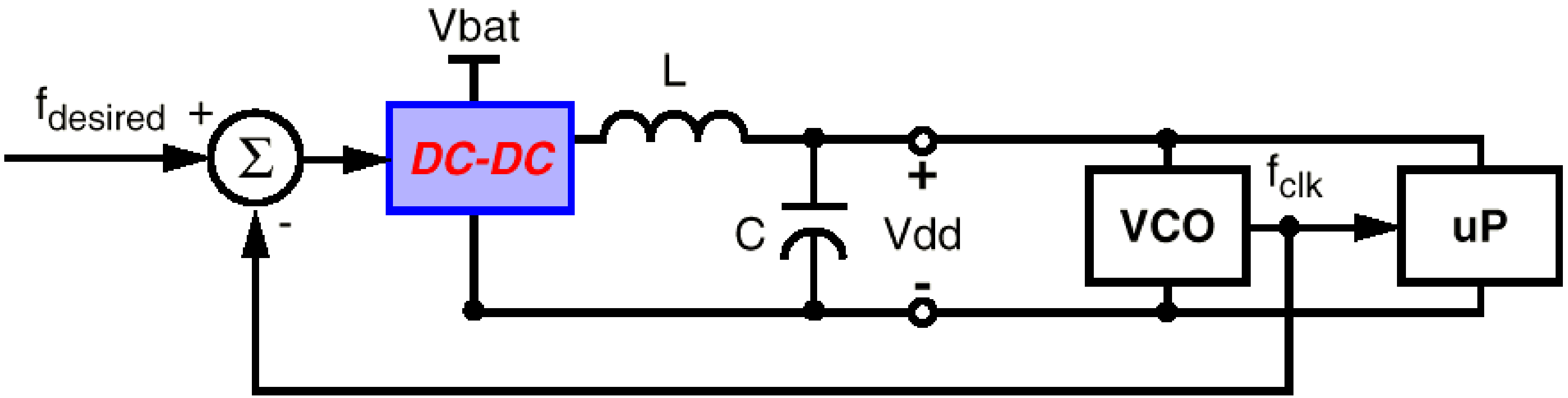


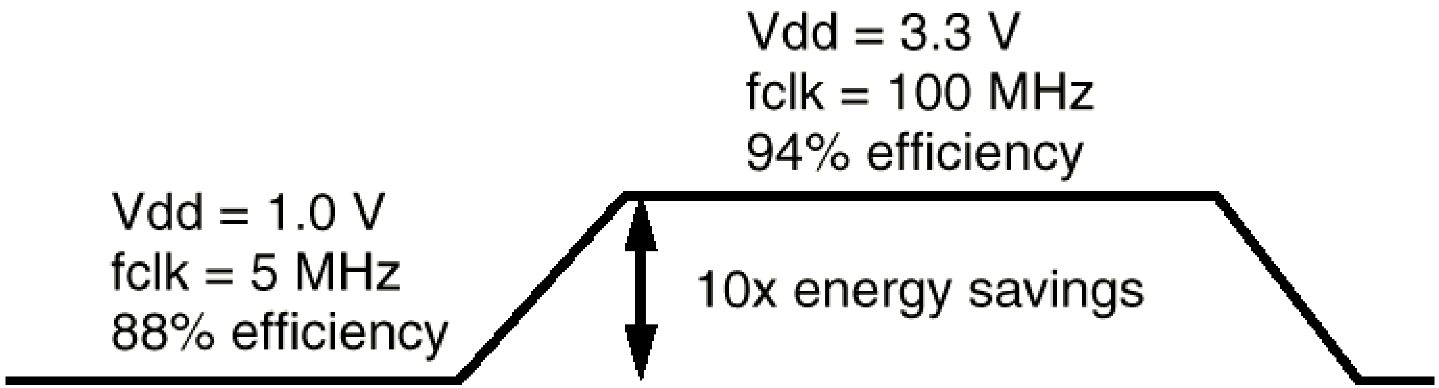


Distributed data-driven control enables globally asynchronous / locally synchronous synchronization

- Avoids overhead of clock distribution
- Enables varying voltages and execution rates
- Modular and scalable

Dynamic Voltage Scaling




Vdd = 3.3 V
fclk = 100 MHz
94% efficiency

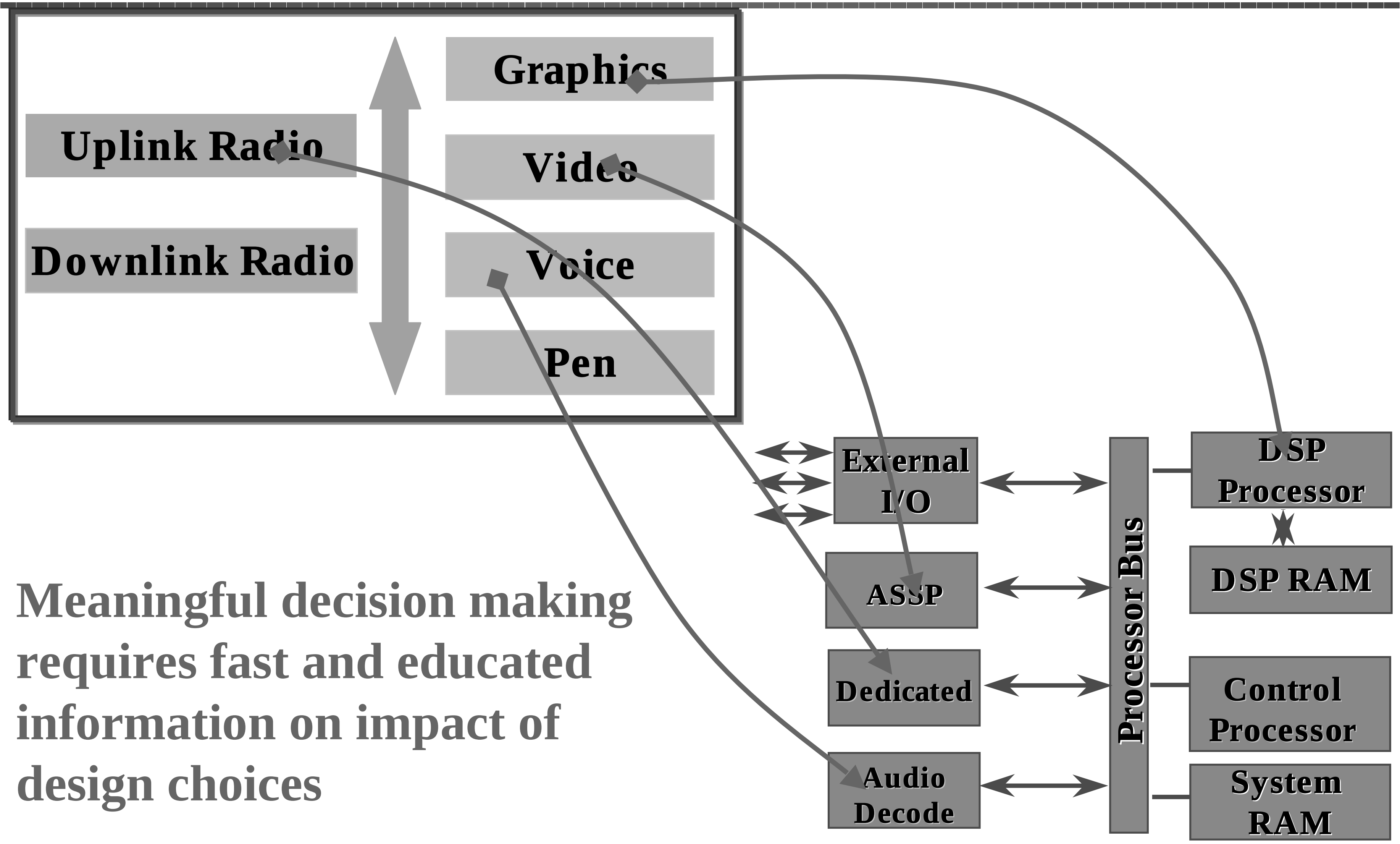
Vdd = 1.0 V
fclk = 5 MHz
88% efficiency

10x energy savings

Design Methodology and Flow

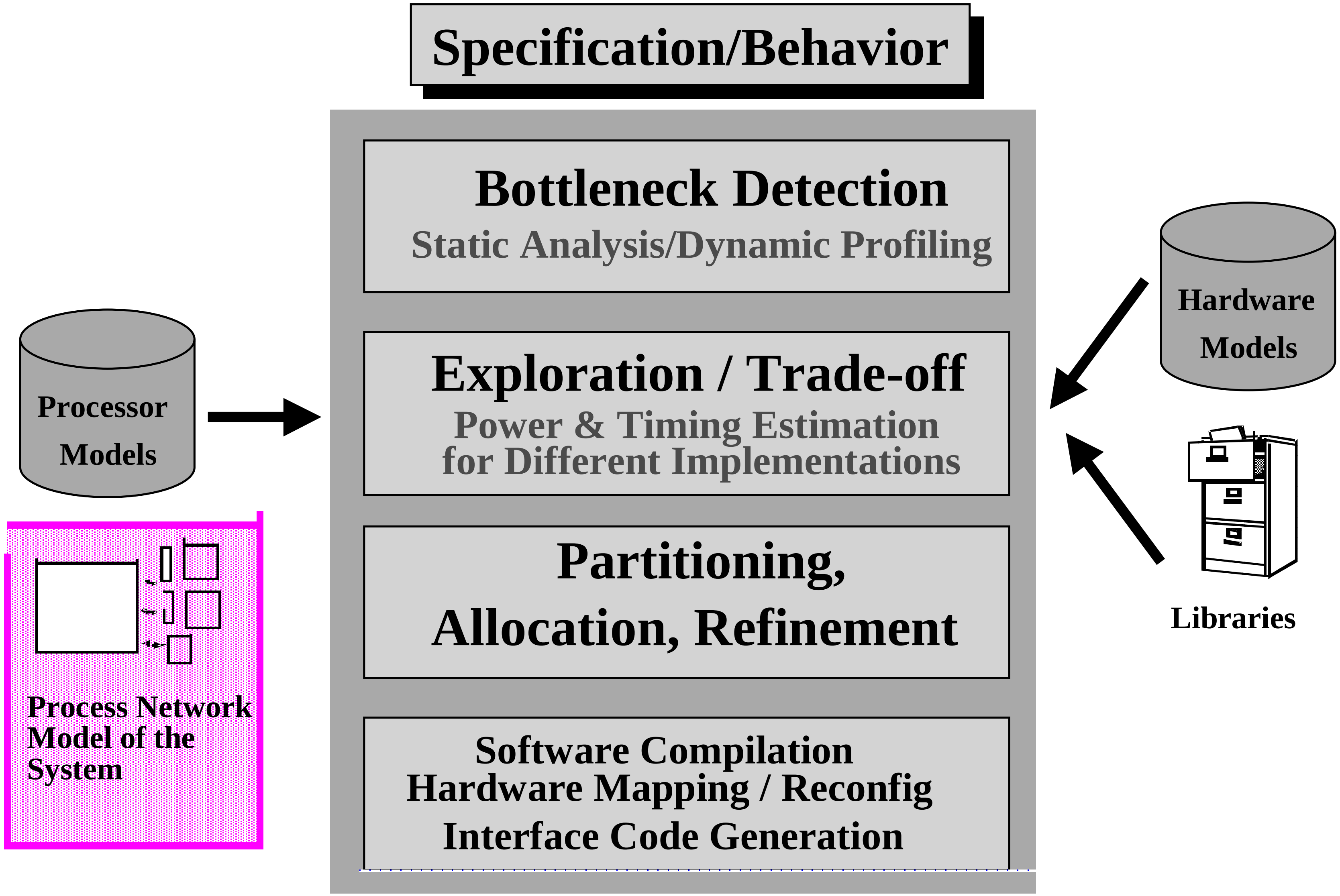
- Requires architecture exploration over heterogeneous implementation fabrics
- Should support refinement and co-design of hardware and software, as well as behavior and architecture
- Should consider all important metrics, or present PDA (Power-Delay-Area) perspective

Behavior-Architecture Co-design



Meaningful decision making requires fast and educated information on impact of design choices

Design Flow for Heterogeneous Architectures



Specification/Behavior

Bottleneck Detection

Static Analysis/Dynamic Profiling

Exploration / Trade-off

Power & Timing Estimation for Different Implementations

Partitioning, Allocation, Refinement

Software Compilation

Hardware Mapping / Reconfig

Interface Code Generation

Processor Models

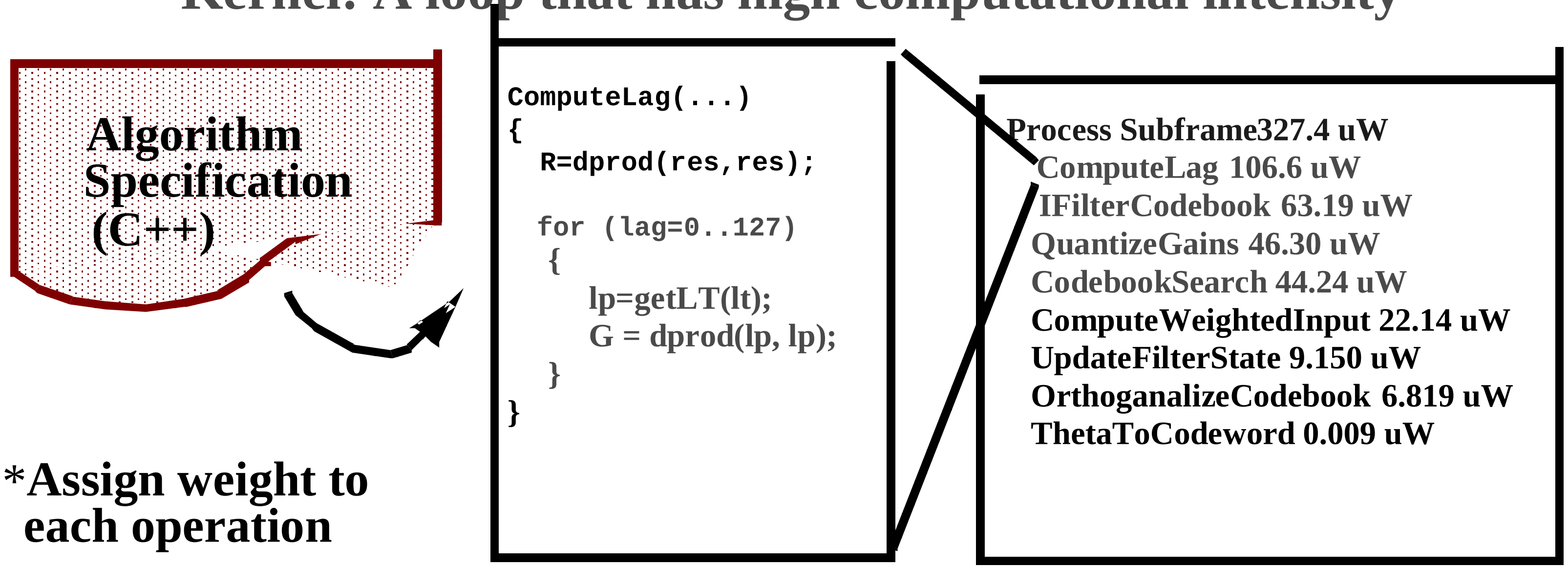
Process Network Model of the System

Hardware Models

Libraries

Bottleneck Detection

Kernel: A loop that has high computational intensity



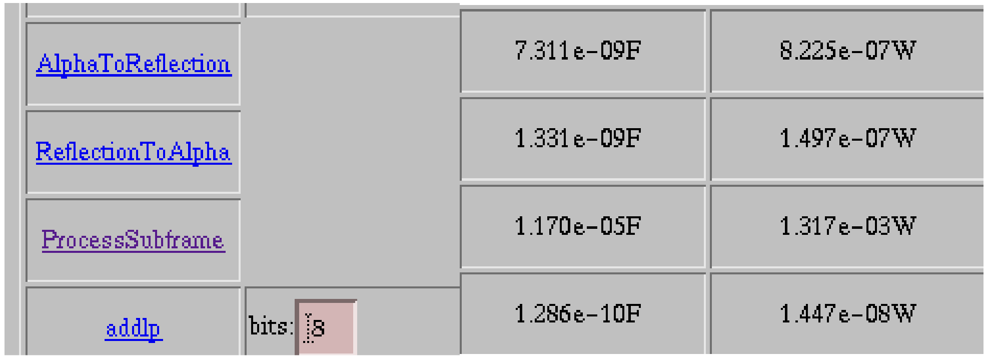
```
ComputeLag(...)
{
   R=dprod(res,res);

   for (lag=0..127)
    {
        lp=getLT(lt);
        G = dprod(lp, lp);
    }
}
```

Process Subframe327.4 uW
ComputeLag 106.6 uW
IFilterCodebook 63.19 uW
QuantizeGains 46.30 uW
CodebookSearch 44.24 uW
ComputeWeightedInput 22.14 uW
UpdateFilterState 9.150 uW
OrthoganalizeCodebook 6.819 uW
ThetaToCodeword 0.009 uW

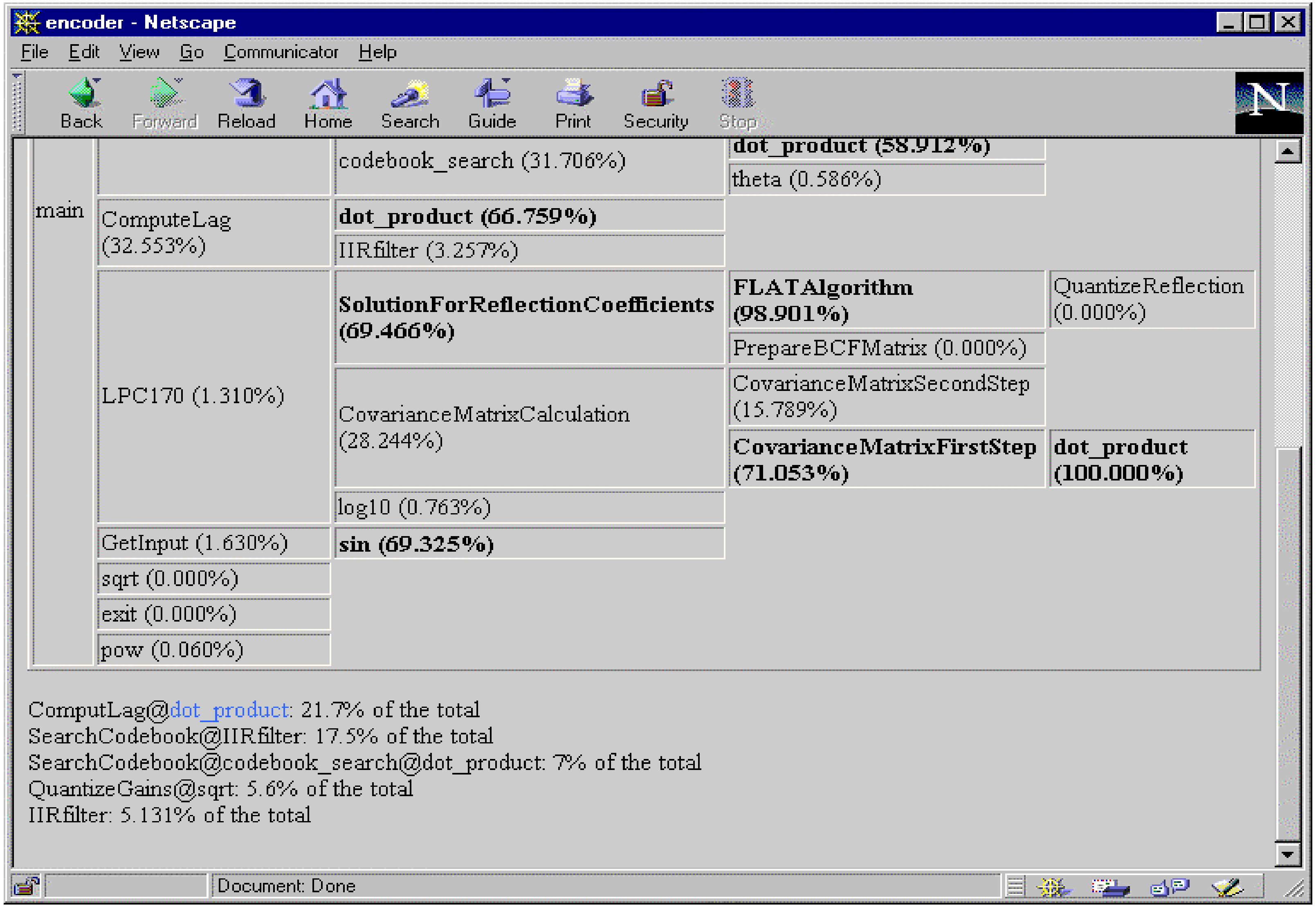
*Assign weight to each operation

*Basic block level Profiling via Quantify, g++, and PowerPlay (UC Berkeley)



AlphaToReflection		7.311e-09F	8.225e-07W
ReflectionToAlpha		1.331e-09F	1.497e-07W
ProcessSubframe		1.170e-05F	1.317e-03W
addlp	bits: 8	1.286e-10F	1.447e-08W

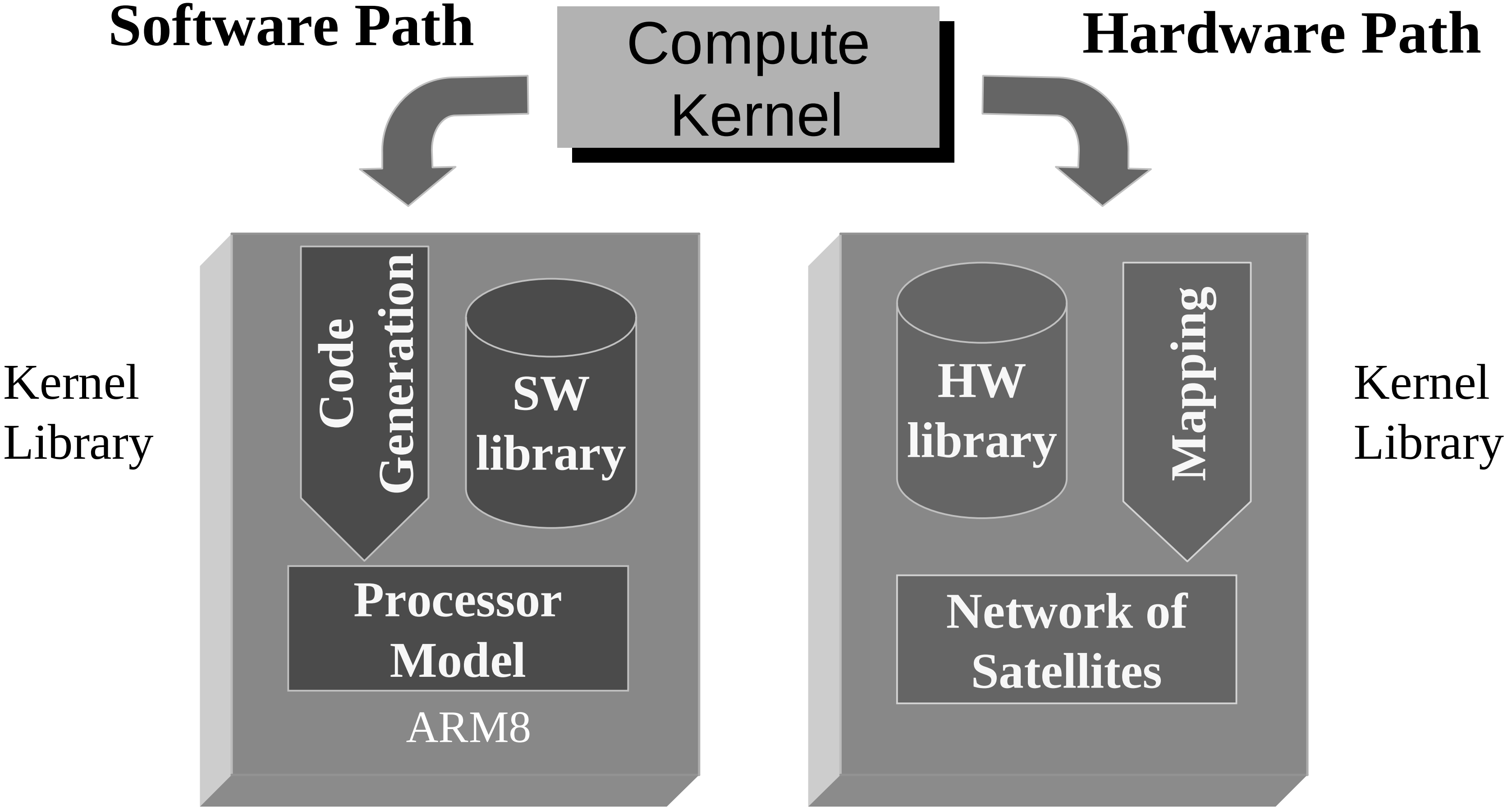
Example: VSELP Speech Coder



main		codebook_search (31.706%)	**dot_product (58.912%)**	
			theta (0.586%)	
	ComputeLag (32.553%)	**dot_product (66.759%)**		
		IIRfilter (3.257%)		
	LPC170 (1.310%)	**SolutionForReflectionCoefficients (69.466%)**	**FLATAlgorithm (98.901%)**	QuantizeReflection (0.000%)
			PrepareBCFMatrix (0.000%)	
		CovarianceMatrixCalculation (28.244%)	CovarianceMatrixSecondStep (15.789%)	
			CovarianceMatrixFirstStep (71.053%)	**dot_product (100.000%)**
		log10 (0.763%)		
	GetInput (1.630%)	**sin (69.325%)**		
	sqrt (0.000%)			
	exit (0.000%)			
	pow (0.060%)			

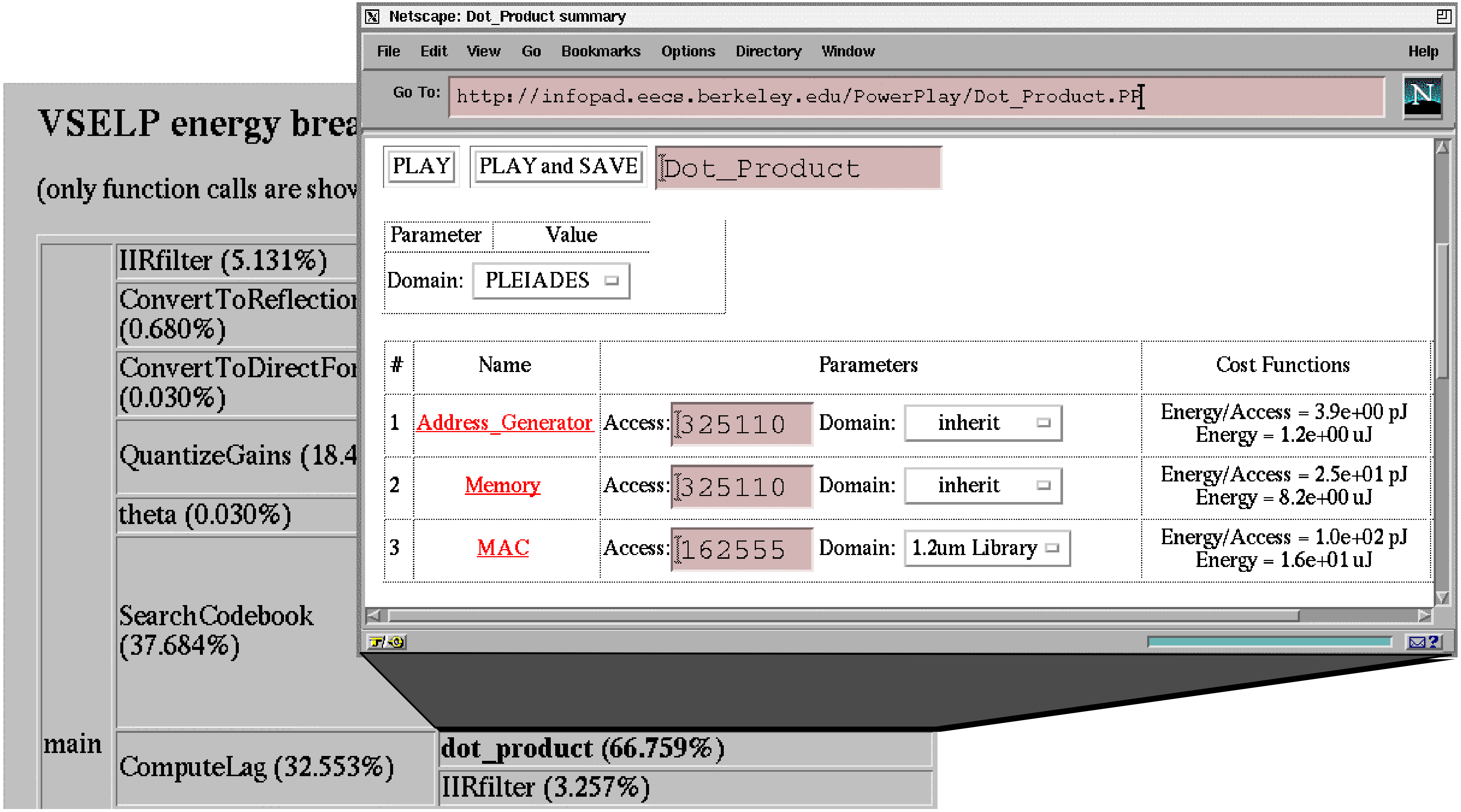
ComputLag@dot_product: 21.7% of the total
SearchCodebook@IIRfilter: 17.5% of the total
SearchCodebook@codebook_search@dot_product: 7% of the total
QuantizeGains@sqrt: 5.6% of the total
IIRfilter: 5.131% of the total

Hardware-Software Exploration



Performance, Power and Area Predictions

Hardware-Software Exploration



VSELP energy brea

(only function calls are shov

main	IIRfilter (5.131%)	
	ConvertToReflection (0.680%)	
	ConvertToDirectFor (0.030%)	
	QuantizeGains (18.4	
	theta (0.030%)	
	SearchCodebook (37.684%)	
	ComputeLag (32.553%)	**dot_product (66.759%)**
		IIRfilter (3.257%)

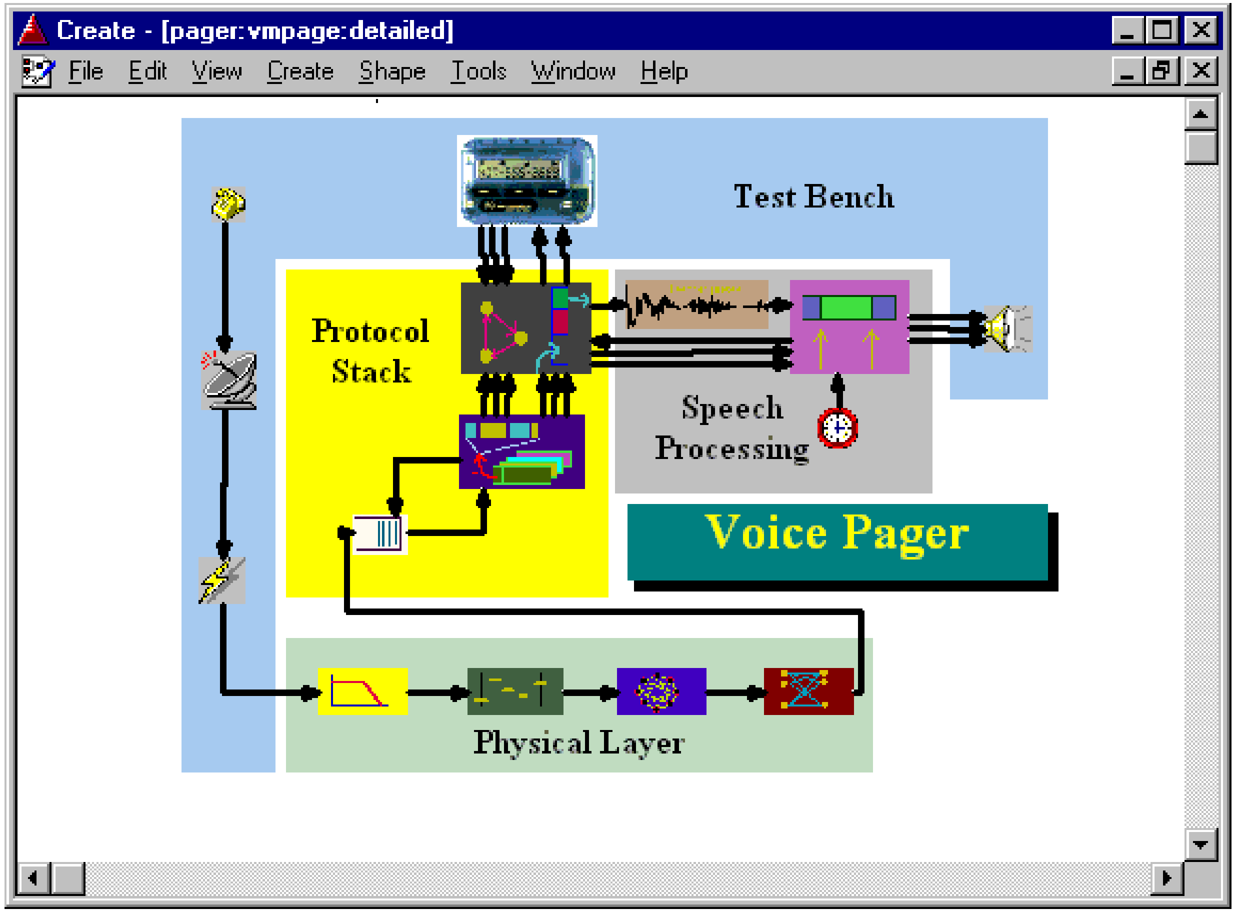
Go To: http://infopad.eecs.berkeley.edu/PowerPlay/Dot_Product.PP

Dot_Product

Parameter	Value
Domain:	PLEIADES

#	Name	Parameters		Cost Functions
1	Address_Generator	Access: 325110	Domain: inherit	Energy/Access = 3.9e+00 pJ Energy = 1.2e+00 uJ
2	Memory	Access: 325110	Domain: inherit	Energy/Access = 2.5e+01 pJ Energy = 8.2e+00 uJ
3	MAC	Access: 162555	Domain: 1.2um Library	Energy/Access = 1.0e+02 pJ Energy = 1.6e+01 uJ

When can we expect to see this from the CAD industry?

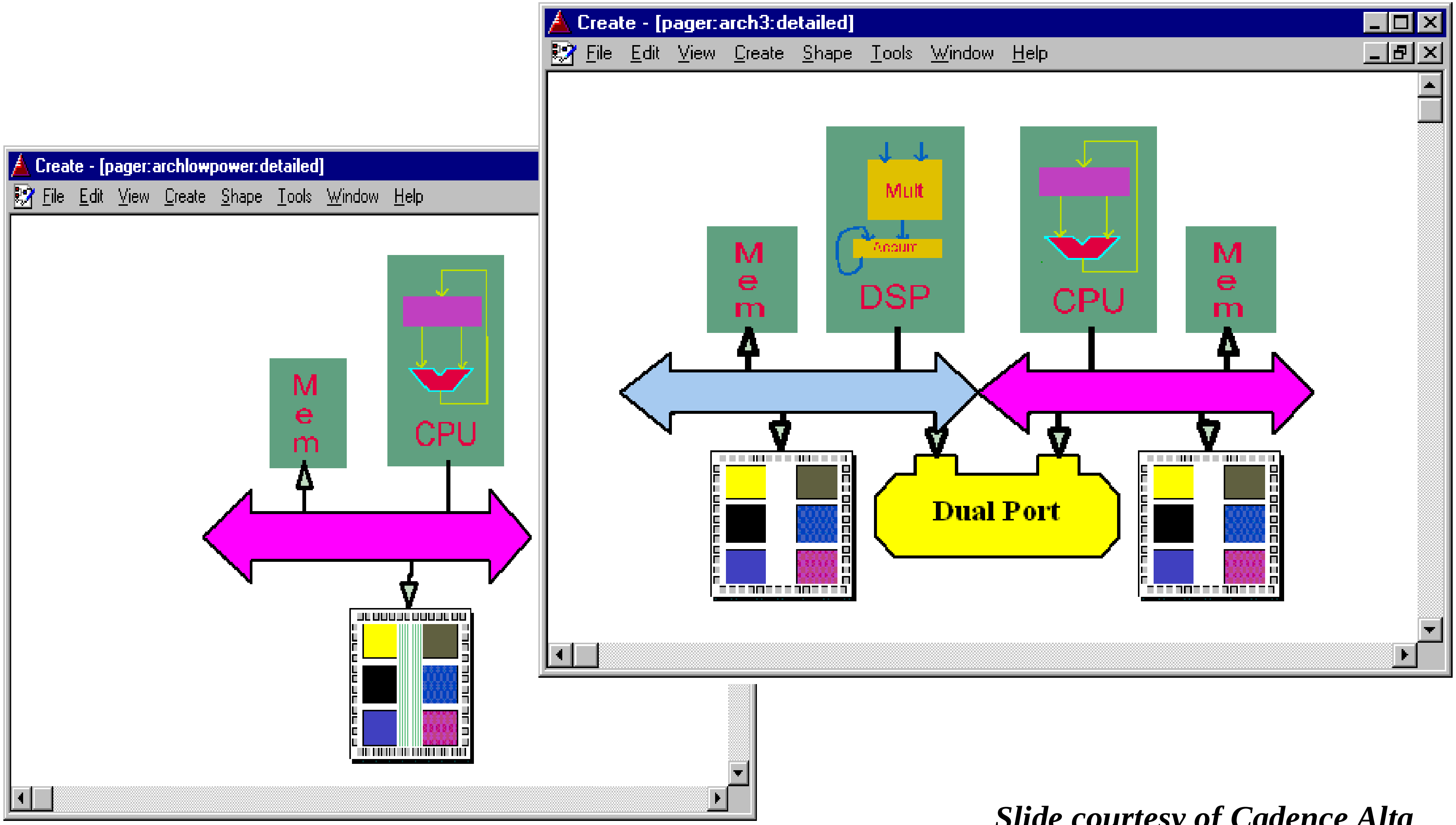


Create - [pager:vmpage:detailed]
File Edit View Create Shape Tools Window Help
Test Bench
Protocol Stack
Speech Processing
Voice Pager
Physical Layer

- Not much activity yet …so don't hold your breath
- Some interesting activities under way that address some of the issues raised
  - » Conceptual design
  - » Hyper-spreadsheets
  - » High-level modeling
- Examples: Co-ware, Alta, Escalade

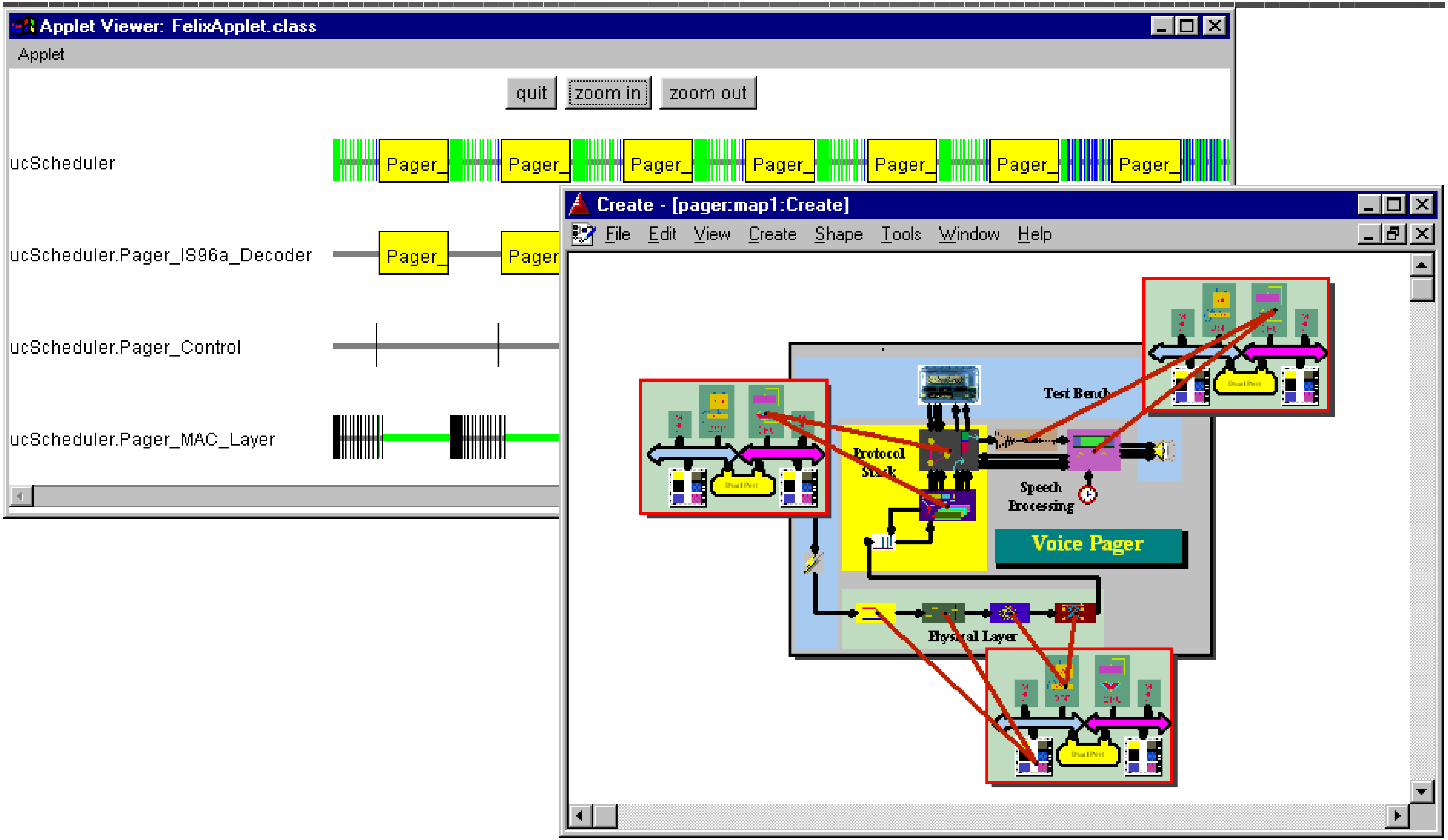
Slide courtesy of Cadence Alta

Architectural Alternatives


Create - [pager:archlowpower:detailed]
File Edit View Create Shape Tools Window Help
Mem
CPU
Create - [pager:arch3:detailed]
File Edit View Create Shape Tools Window Help
Mem
Mult
Accum
DSP
CPU
Mem
Dual Port

Slide courtesy of Cadence Alta

Impact of Mapping on Performance



Summary

- Reconfiguration — A new paradigm in computation: a greater role for interconnect!
- Matching granularity in computation and architecture opens opportunities for high performance / low energy computing
- Heterogeneous architectures are probable choice for system on-a-chip, yet pose important challenges in terms of applicability and flexibility