Preliminary Performance Results of High Performance Embedded Computing (HPEC) on KASSPER Algorithms

Matthew French

mfrench@isi.edu

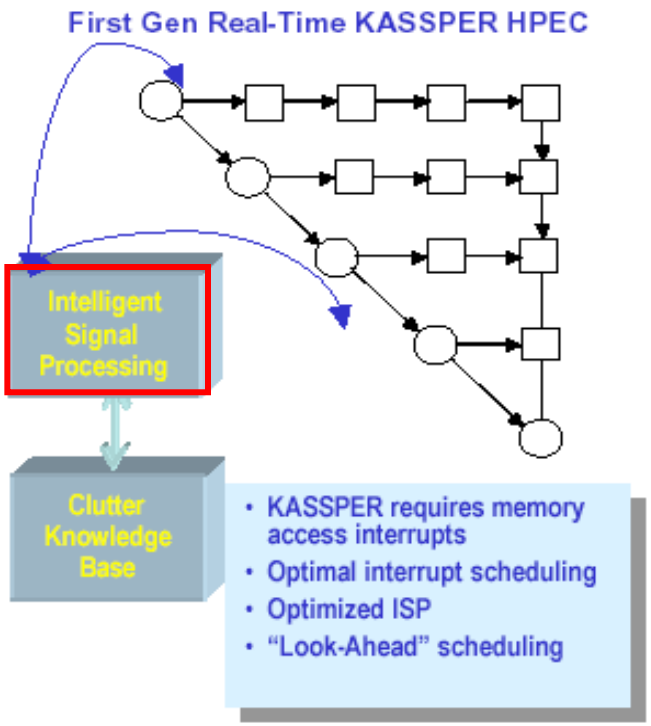
University of Southern California

Information Sciences Institute

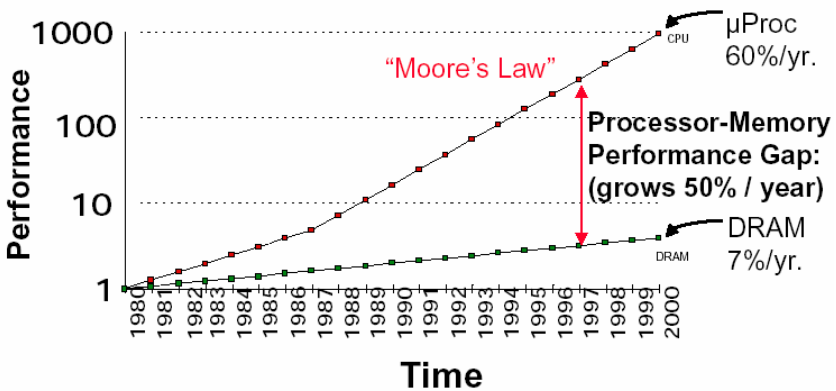
April 7th, 2004

Motivation

- **KASSPER algorithms blend Knowledge processing with traditional signal processing (Threaded and Streaming)**
- **Traditional Signal Processing**
  - **Streaming**
  - **FLOPs**
- **Knowledge Processing**
  - **Database Access Rates**
  - **Large Memory Structures**
- **New Set of Computer Architecture Parameters Stressed**
  - **Memory Latency**
  - **I/O Throughput**
  - **Multi-threaded Application Performance**
  - **Data Locality**



Processor-DRAM Gap (latency)

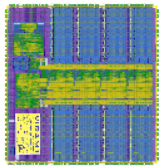


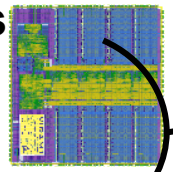
HPEC role in KASSPER

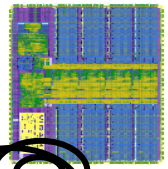
- KASSPER Architecture is a system level problem
- Correct choice of micro-architecture atomic unit removes bottlenecks and simplifies system
- HPEC micro-architectures will drastically change system level requirements
  - SWEPT
  - Interconnect
  - Type / Size of Memory Banks

System A

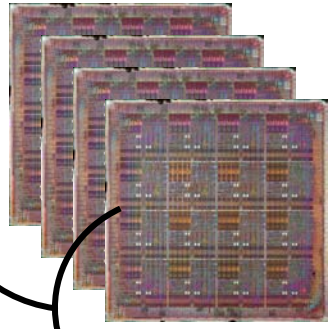
- High BW Interconnect
- No additional memory
- 1000's of processors







System B



- Interconnect sufficient
- Additional distributed memory
- 100's of processors

Different Micro-Architecture = Different System Trade-offs

Candidate Processing Units

PowerPC

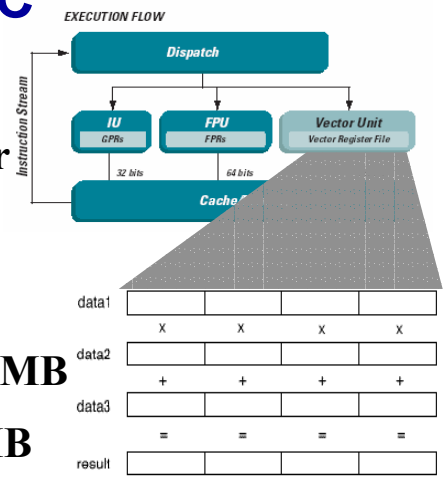
RISC Architecture with Altivec Co-Processor

500 MHz

4 GFLOPS Peak (MAC)

SDRAM: 125 MHz, 256 MB

L2 cache: 250 MHz, 2 MB



V-IRAM

Processor In Memory (PIM) Architecture

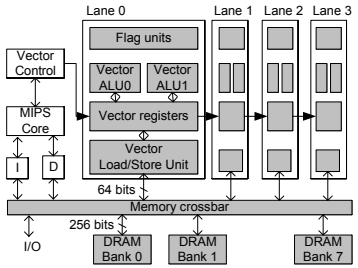
200 MHz; 13MB DRAM

6.4 16-bit GOPS, 1.6 GFLOPS

4 float ALUs; 8 32bit int ALUs; 16 16bit ALUs

12.8 GB/s peak memory access

http://iram.cs.berkeley.edu/



Imagine

Streaming Architecture

300 MHz, VLIW SIMD machine

28 16-bit GOPS, 14 GFLOPS

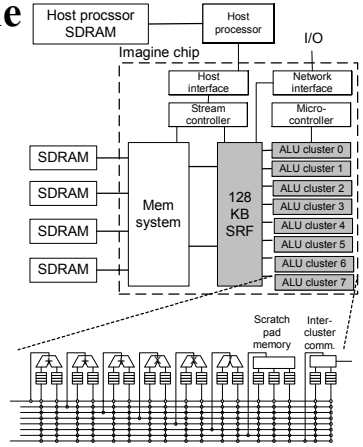
128 kB Streaming Register File

8 ALU Clusters

- 6 ALUs / cluster
- 84-95% ALU utilization typical

Streaming Memory Buffers

http://cva.stanford.edu/



Raw

Tiled Architecture

16 tiles of MIPS R4000 @ 425 MHz (chip)

- 6.8 GOPS or GFLOPS

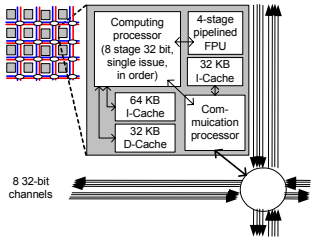
4 Communication Networks

- 2 Static Networks,
  - 1 cycle throughput
  - 3 cycle latency
  - 54.4 GB / sec
- 2 Dynamic Networks

14 External Ports (I/O or DRAM)

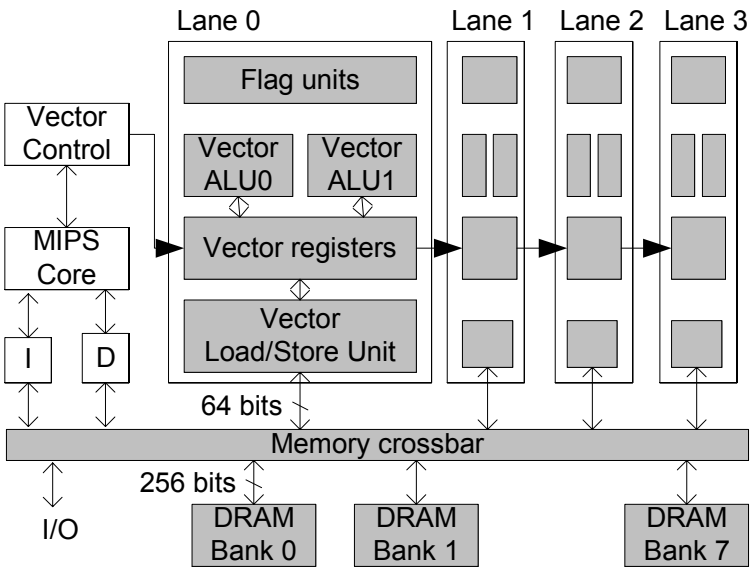
- 47.6 GB/sec

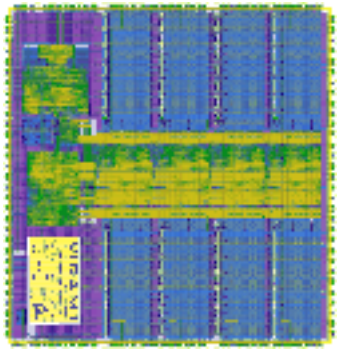
http://www.cag.lcs.mit.edu/raw/



Vector Intelligent RAM (VIRAM, Berkeley)

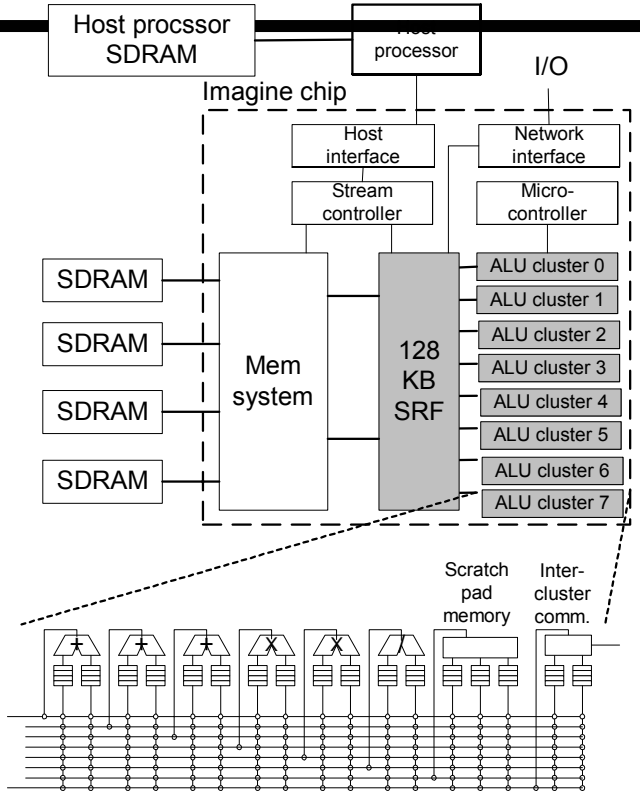
- Merge DRAM with Vector Processor
- mixed logic-DRAM CMOS process
- Scalar MIPS processor core
- 6.4 16-bit GOPS, 1.6 GFLOPS
  - 4 float ALUs; 8 32bit int ALUs; 16 16bit ALUs
- 12.8 GB/s peak memory access
- 13 MB DRAM
- 15 x 18 mm; IBM Foundry
- Chips fabbed in Q1 ’03
- C/C++ w/ pragmas, ASM; Cray PDGCS compiler
- Can add additional external DRAM
- http://iram.cs.berkeley.edu/





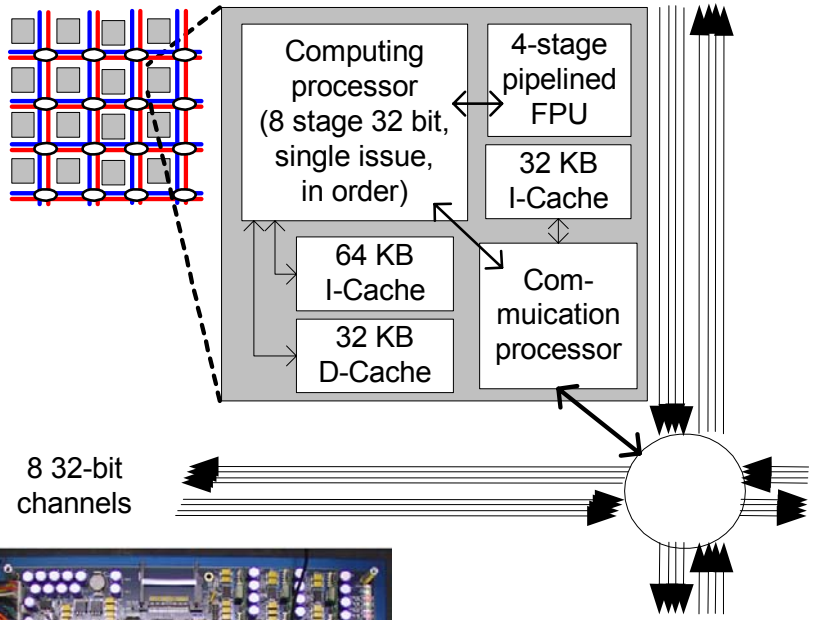
Imagine Streaming Processor (Stanford)

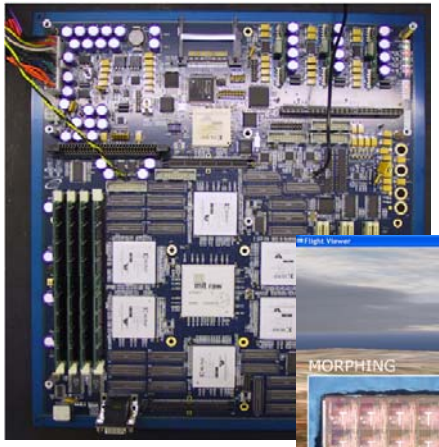
- 300 MHz, VLIW SIMD machine
- 28 16-bit GOPS, 14 GFLOPS
- 128 kB Streaming Register File
- 8 ALU Clusters
  - 6 ALUs / cluster
  - 84-95% ALU utilization typical in multimedia
  - 256 x 32 bit local register file
- Streaming Memory Buffers
  - re-order DRAM accesses
  - expose data locality
  - ALU Intra-cluster BW - 435 GB/sec
  - DRAM BW - 2.1 GB/sec
- 16 x 16 mm; TI Foundry
- StreamC & KernelC programming languages
- Network interface for scalability
- Chip and board functional;
- http://cva.stanford.edu/

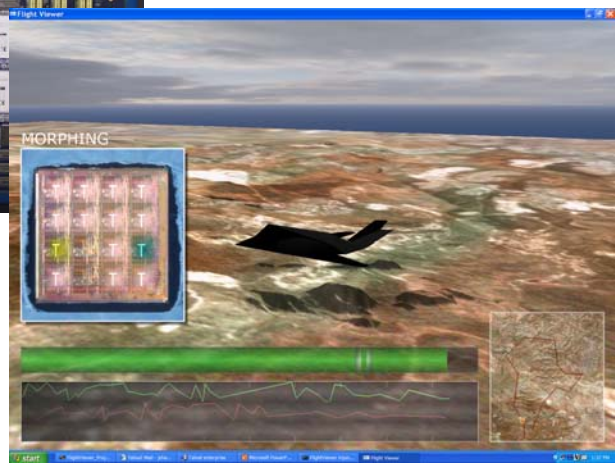


Raw (MIT)

- 16 tiles of MIPS R4000 @ 425 MHz
  - Hardware limited to 300 MHz
  - 6.8 GOPS or GFLOPS
- 4 Communication Networks
  - 2 Static Networks,
    - 1 cycle throughput
    - 3 cycle latency
    - 54.4 GB / sec
  - 2 Dynamic Networks
- 14 External Ports (I/O or DRAM)
  - 47.6 GB/sec
- C and ASM; gcc based compiler
- 18.2 x 18.2 mm; IBM Foundry
- Fully scalable architecture
- Demonstration on Hardware at DARPATech 2004
- http://www.cag.lcs.mit.edu/raw/







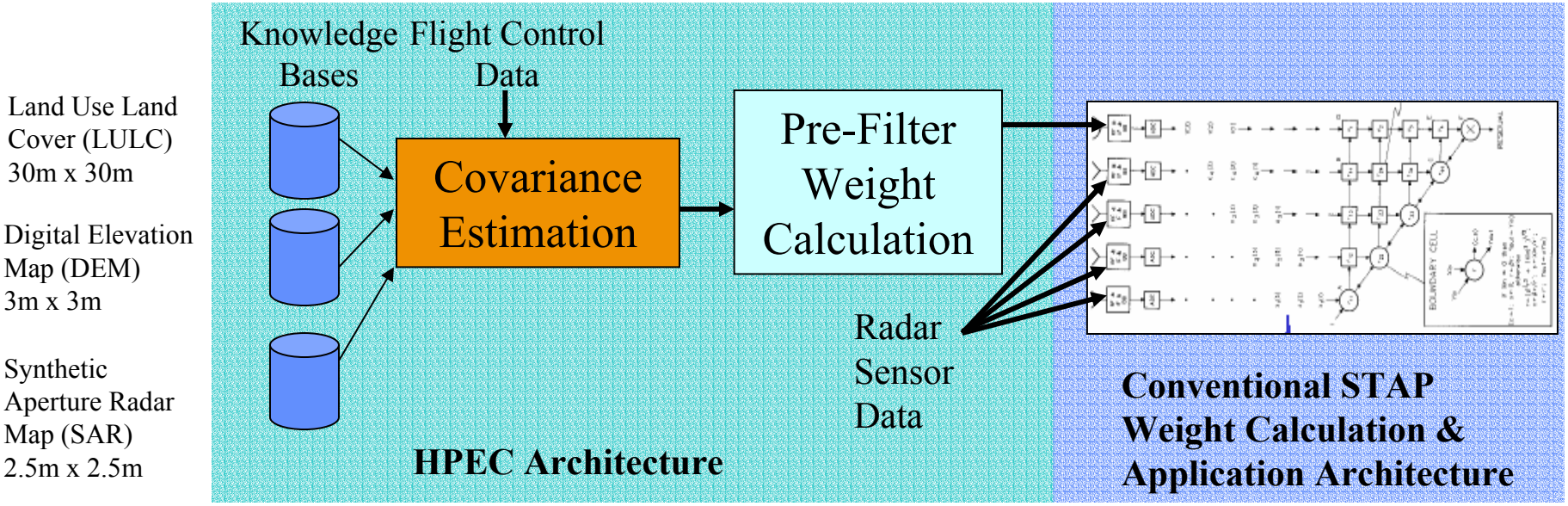
Technical Approach

- Analyze System Requirements
  - Explore Memory vs Computation Trade-offs
- Define Memory Requirements
  - Understand Databases
    - Total Size
    - Element Resolution
    - What portion to access per CPI, per pulse?
  - Derive Memory Size, Bandwidth Parameters
- Define Necessary Computations
  - Derive Computational Throughput Parameters
  - Validate Computations
- Consider Look-Ahead Scheduling Kernel
  - Define Kernel for each Architecture
- Map Kernels, Optimize, & Measure Performance

Pre-Filter Algorithm – Data Domain

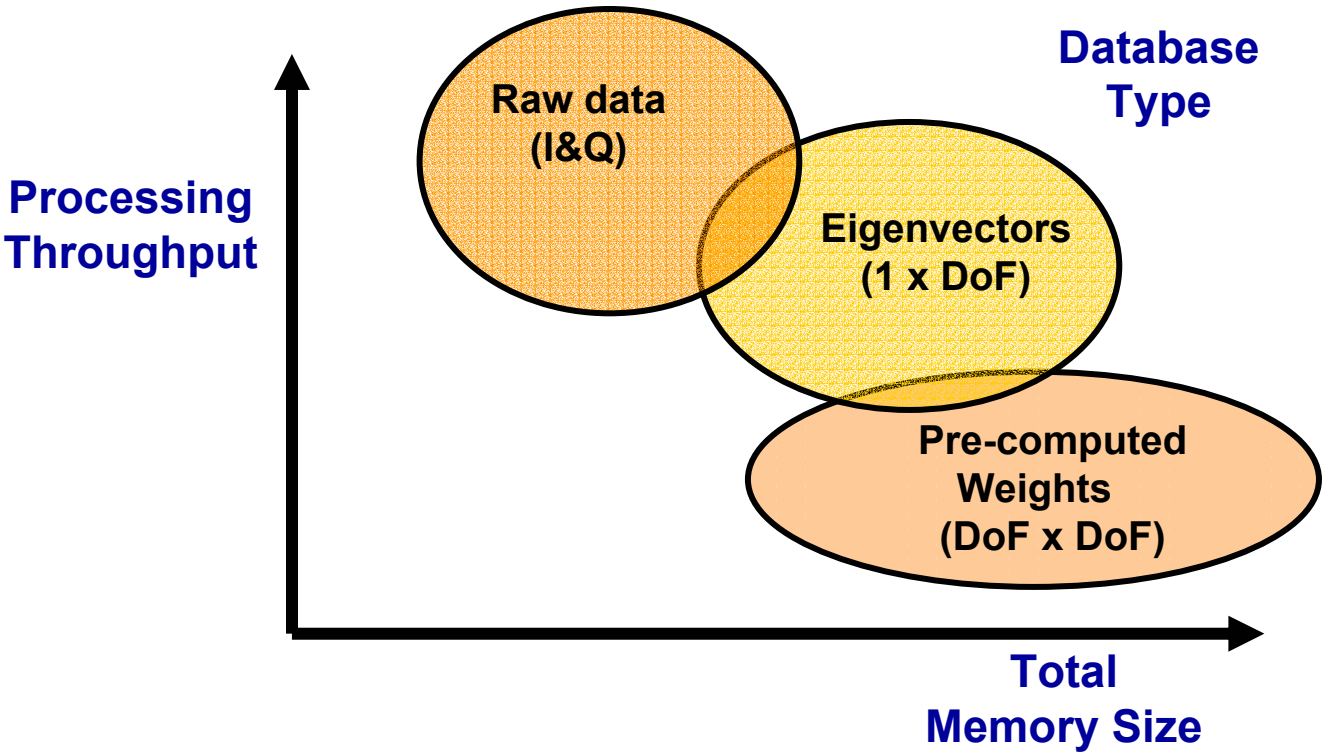
$$\mathbf{w} = \frac{(\mathbf{R}_{xx} + \beta_d \mathbf{R}_c + \beta_L \mathbf{I})^{-1}\mathbf{v}}{\mathbf{v}^H(\mathbf{R}_{xx} + \beta_d \mathbf{R}_c + \beta_L \mathbf{I})^{-1}\mathbf{v}} = \frac{(\mathbf{R}_{xx} + \mathbf{Q})^{-1}\mathbf{v}}{\mathbf{v}^H(\mathbf{R}_{xx} + \mathbf{Q})^{-1}\mathbf{v}}$$

$$\mathbf{L}\mathbf{L}^H = \begin{bmatrix}\mathbf{x} & \mathbf{C}\end{bmatrix}\begin{bmatrix}\mathbf{x}^H \\ \mathbf{C}^H\end{bmatrix} = \mathbf{R}_{xx} + \mathbf{Q}$$



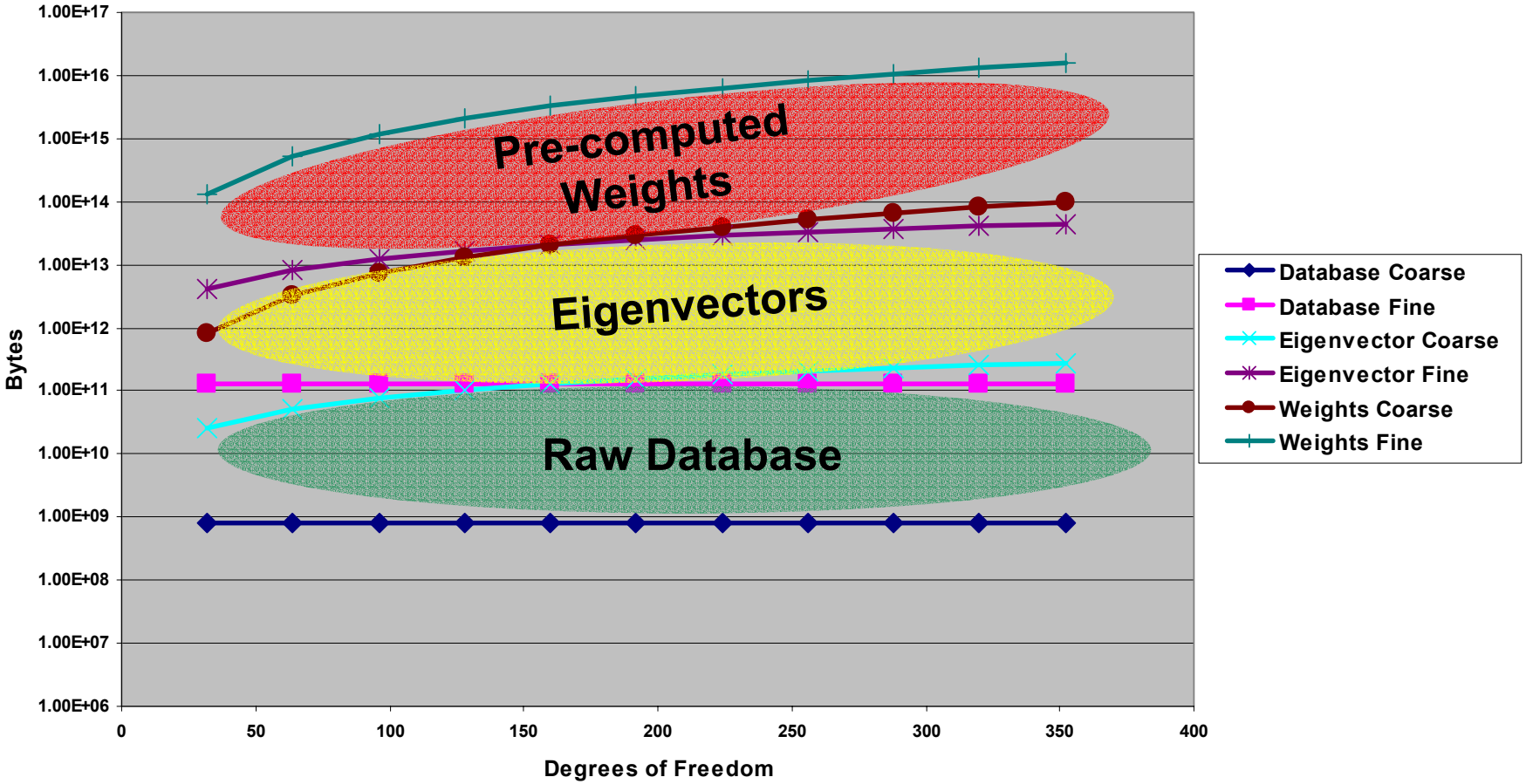
- **Covariance Estimation forms $\mathbf{Q} = \beta_d \mathbf{R}_c + \beta_L \mathbf{I}$**
- **Prefiltering Stage Forms $\mathbf{C} = \mathbf{Q}^{1/2}$**
- **Conventional Processing Performs Color Loaded QR Decomposition**

System Requirements

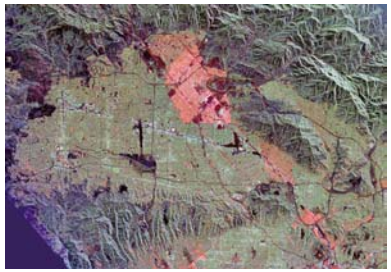


- **Memory and Computation Requirements Driven by Database Type**
  - Weight Database
  - Original SAR, DEM, LULC, etc
  - Eigenvector or some intermediary form
- **Further Analysis to Label Axis, Determine Shapes, Check Feasibility…**

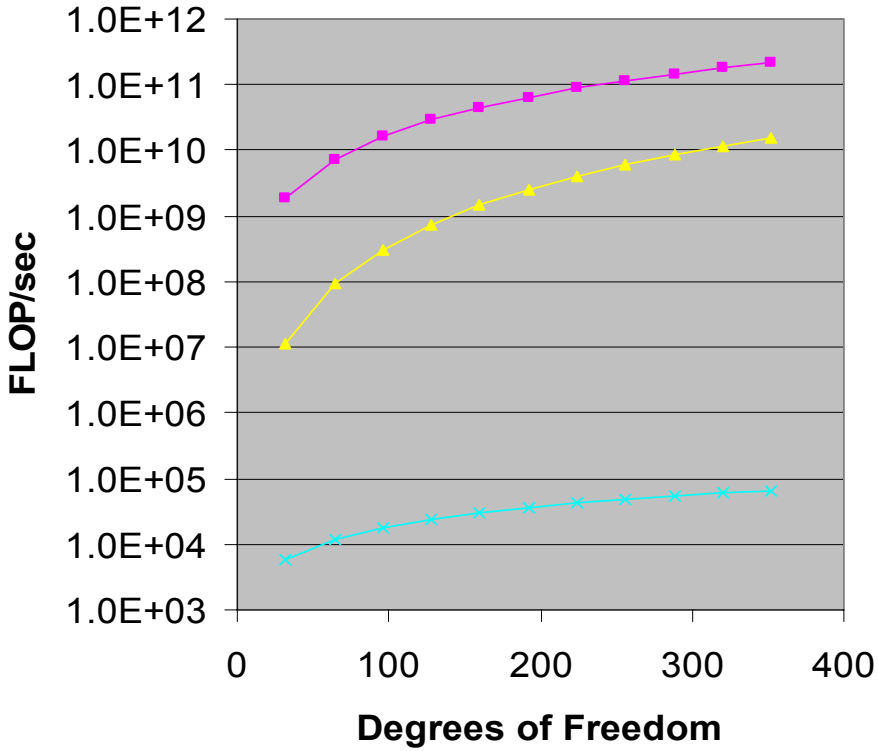
Memory Requirements



- **Region of Interest Size: 100,000 sq km**
  - SAR (2.5m x 2.5m): 1.6 x 10^{10} cells - Fine
  - LULC (30m x 30m): 1x 10^{8} cells - Coarse



Throughput Requirements

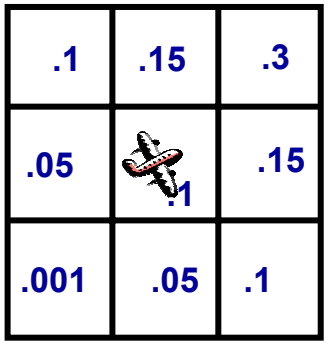


Kernel	Database	Eigenvectors	Weights
Clutter Cov Formation	X		
SVD Decomposition	X		
R_{known} Formation	X	X	
Cholesky Decomposition	X	X	
Database Indexing	X	X	X

- **Database vs Eigenvectors**
  - 10x More Throughput FLOPS
  - 100x Less Memory Required

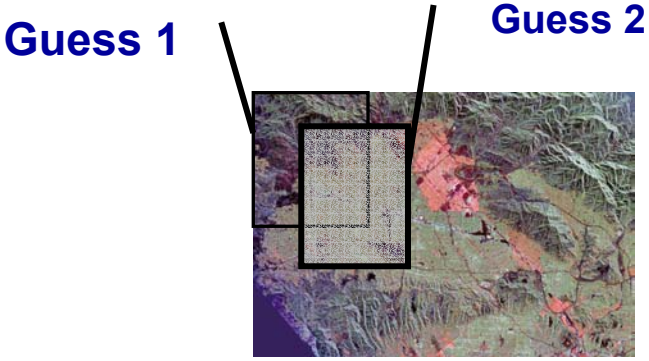
Look Ahead Observations

- **Look-Ahead Adds Computational Complexity**
  - Simple Case – move one cell / CPI
    - 100 m/s platform speed
    - 2.5 m x 2.5 m cell resolution
    - =25 ms / cell; CPI = 16 ms
  - ~9 Guesses for 100% coverage, ~9x Operations
  - Prioritize by Probability



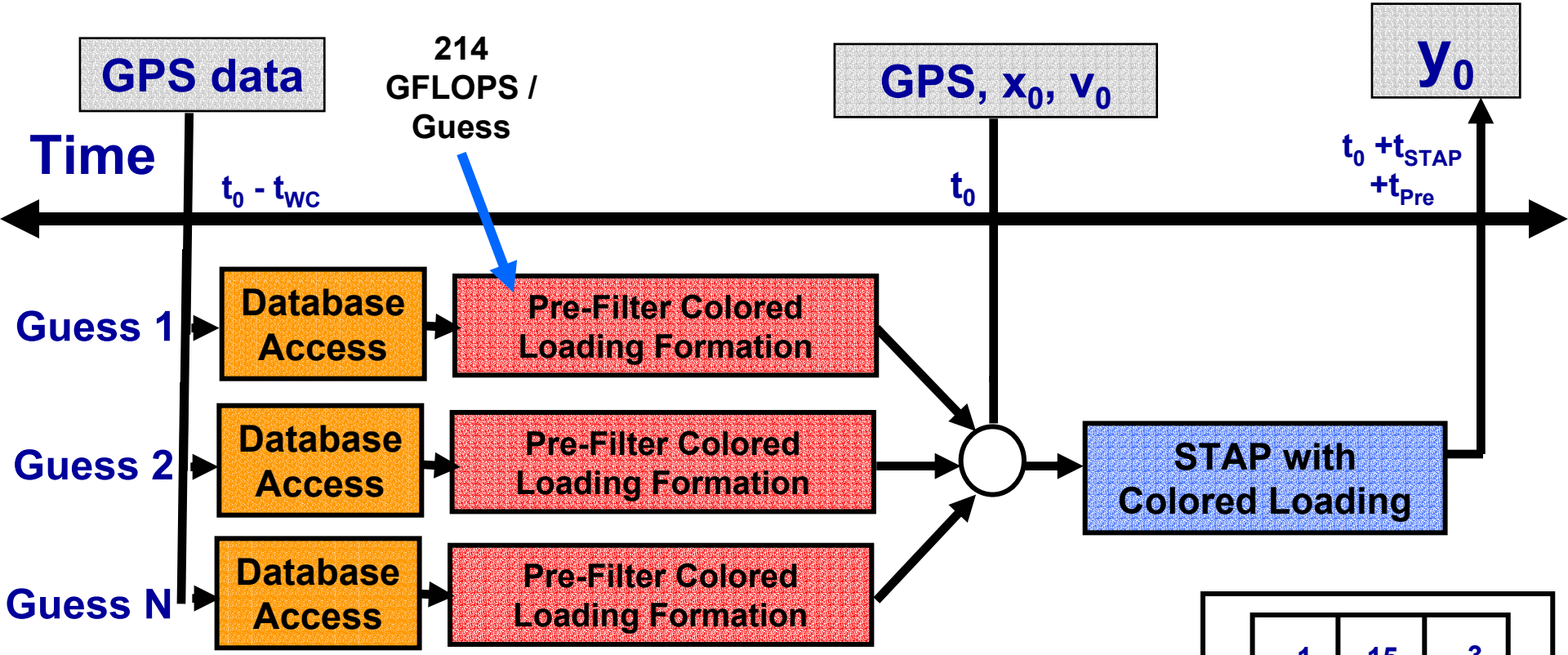
Future Cell Location Probability

- **Exploit Data Locality**
  - Translates to less load and stores
  - Partial results re-use?



- **Kernels Highly Architecture Dependant**
  - Leverage Memory Hierarchy Strengths
- **Memory Bandwidth Requirements**
  - Driven by Look Ahead Computation Techniques
- **Potentially Further Performance Gains by Using HPEC Technology**

Look Ahead Scheduling Timeline



.1	.15	.3
.05	.1	.15
.001	.05	.1

Future Cell Location Probability

- Pre-Compute Multiple Weight Sets (Guesses)
  - Choose best Guess at time t_0
- Computational Load Increased
- Database Accesses Likely to Have High Overlap

Data Domain Computational Analysis

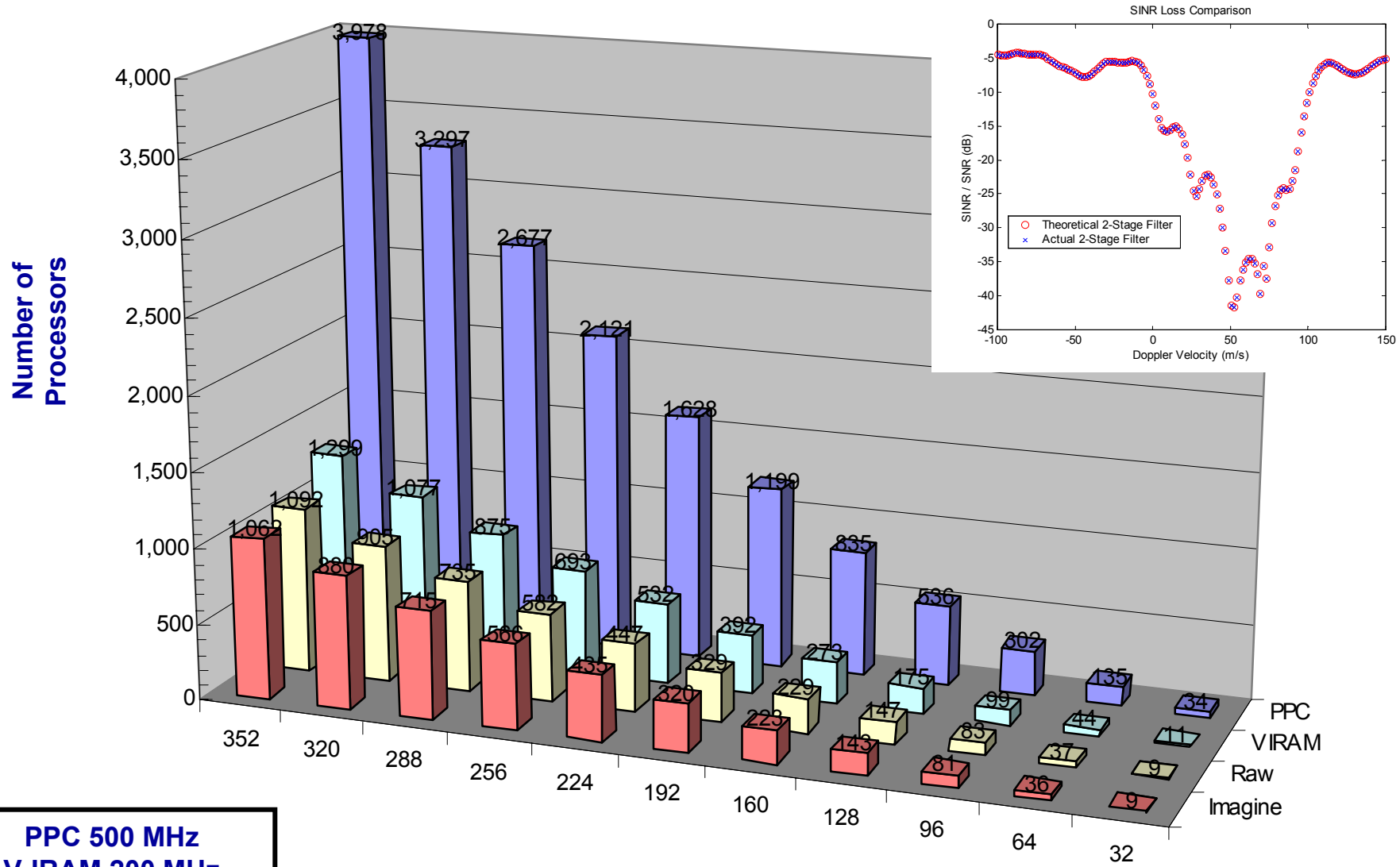
Operation	Order of Op	FLOPs/CPI	FLOPs/sec One Guess	Prediction FLOPs/sec
Clutter cov formation	$O(N^2*M)$	1.8E9	1.1E11	8.8E11
SVD decompostion	$O(N^2*M)$	1.5E9	9.4E10	7.5E11
R_{known} Formation	$O(N^3)$	1.0E8	6.3E9	5.0E10
Cholesky Decomposition	$O(N^3)$	5.83E7	3.6E9	2.9E10
Total Weight Calculation		**3.46E9**	**2.14E11**	**1.71E12**

Using KASSPER Data Set 1 Parameters

M = Database vector length = 1800; N = DOF = 352;
R = Range Gates = 1000; CPI = 16ms

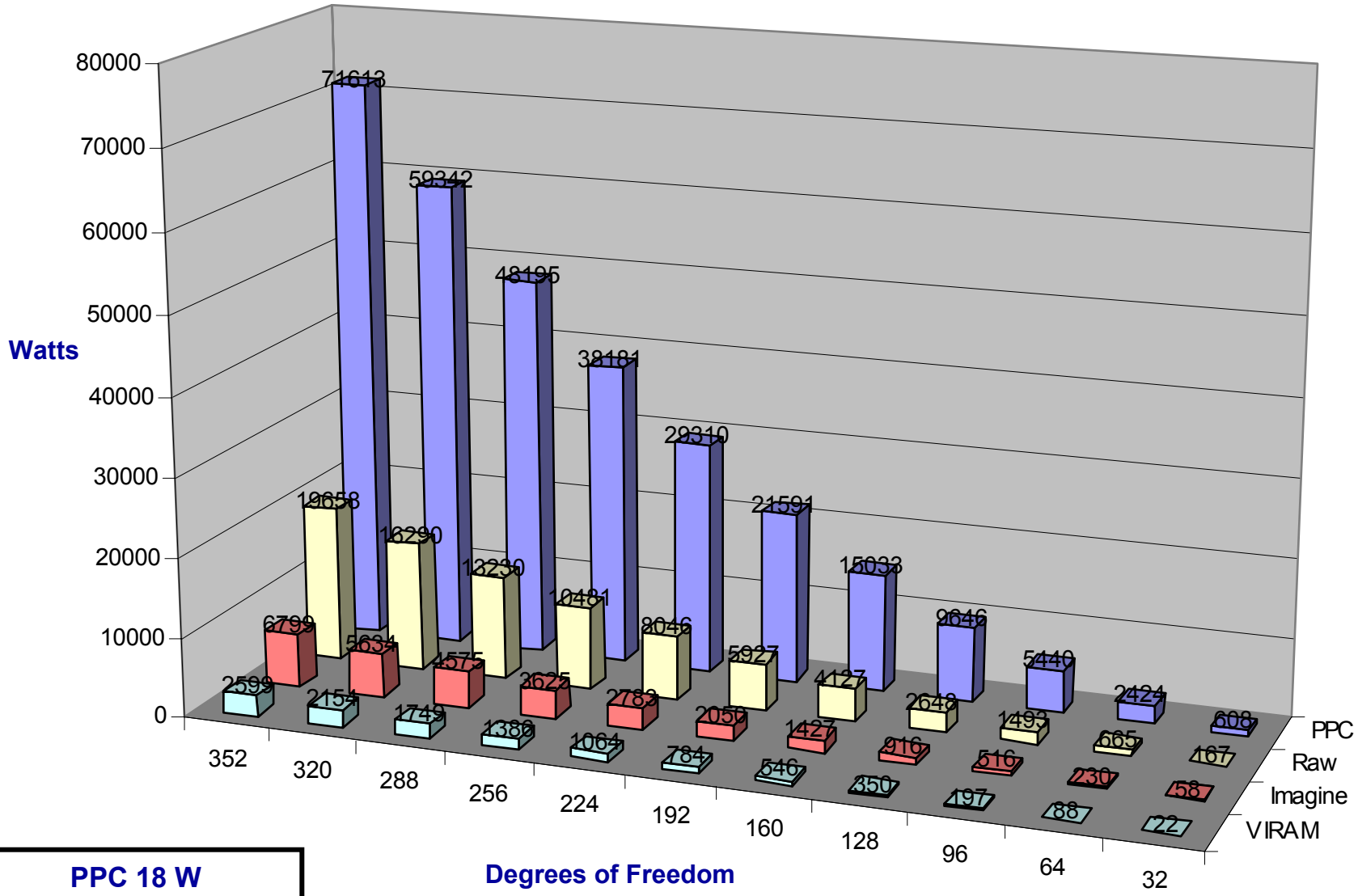
- **Prediction – Assuming 9 Future Cell Location Guesses**
  - **Platform speed = 100 m/s; Cell resolution = 2.5m x 2.5m; ≈ 1 Cell / CPI**
  - **9 Locations in a 2-D grid**
  - **More if consider 3-D case**
- **1.71 Tera Flops/sec Minimum Throughput**

Pre-Filter Prediction System Sizing



PPC 500 MHz
V-IRAM 200 MHz
Raw 300 MHz
Imagine 300 MHz

Pre-Filter Prediction Power



PPC 18 W
V-IRAM 2 W
Raw 18 W
Imagine 6.4 W

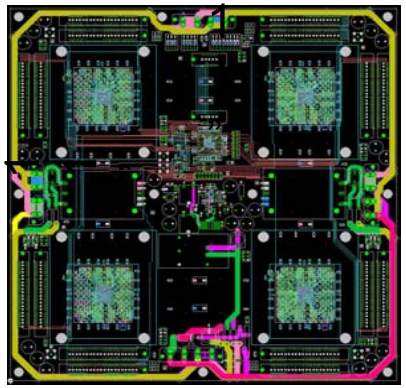
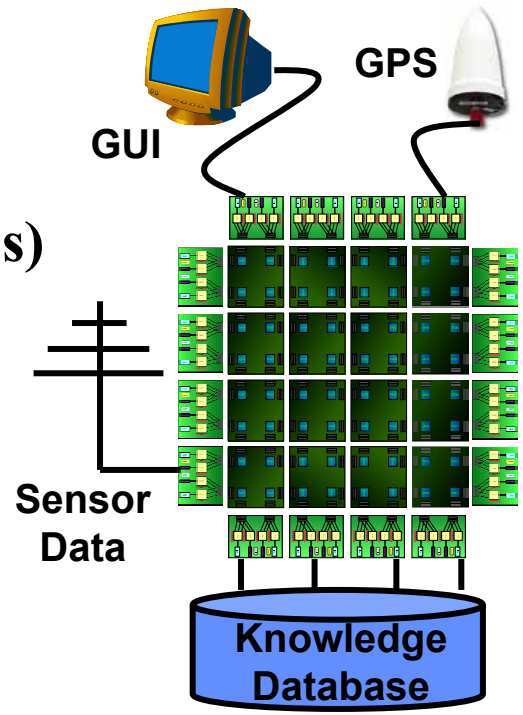
Preliminary Architecture Comparison

Architecture	System Maturity	Kernel Processing Throughput	Scalability	Versatility	Power	Ease of Programming	Level of Support	Ongoing DARPA Funding
PowerPC	Commer-cial	1	Medium	High	20W typical	Easy to Medium	High	NA
Raw	High	4x PPC	High	High	18W typical	Medium	Medium	Yes
Imagine	Medium	4x PPC	More Research Needed	Medium	6.4W typical	Medium to Difficult	Medium	No
V-IRAM	Medium	3x PPC	More Research Needed	Medium	2W typical	Difficult	Low	No

- Size and weight are nearly identical for all architectures
- Raw
  - Robust performance across all system metrics
  - MAC Instruction Not Present in Raw
    - ~2x improvement for all kernels
- Research Chips
  - Small Team of PhD Students vs Industry Professionals
  - Standard Cells vs Custom ASIC Logic
- Detailed Results Published in IEEE Radar 2004

Future KASSPER Direction

- Multi-Processor Hardware in Development
- 20 DOF system (4 Raw Chips)
  - Key Aspects of System
  - Scale up kernels
    - Close to full size
    - Focus on Critical Kernels – Flops and Mem interface
    - Will yield more accurate performance estimates
  - Partial use of KASSPER Data Set #2



- 100 DOF Proto-type System (64 Raw Chips)
  - Baseline with:
    - Flight prediction algorithm
    - GUI
    - GPS mock interface
    - STAP back-end emulation