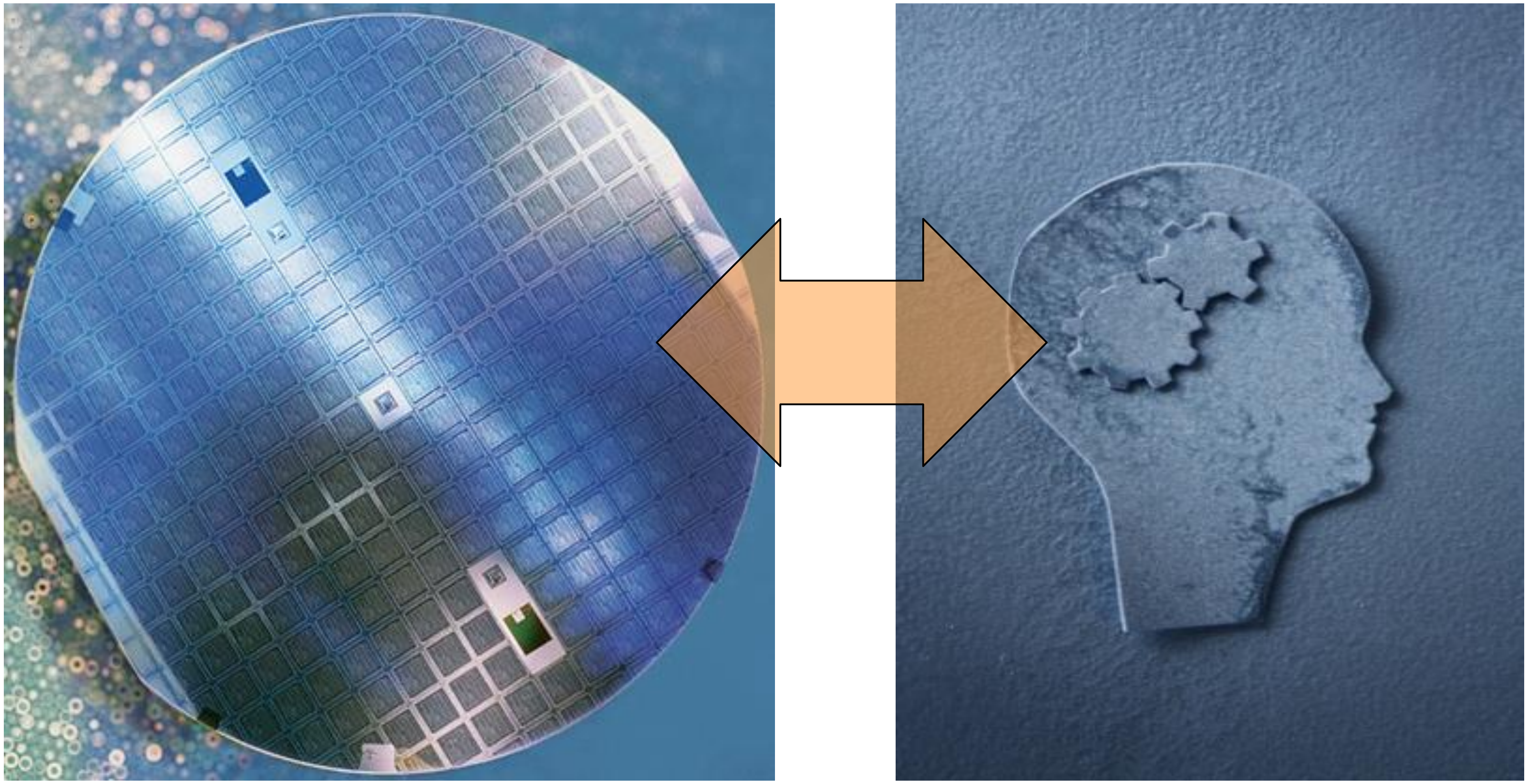
Custom Hardware Accelerators for Statistical Inference for Machine Learning

Rob A. Rutenbar
Bliss Professor & Head
Department of Computer Science
University of Illinois at Urbana-Champaign

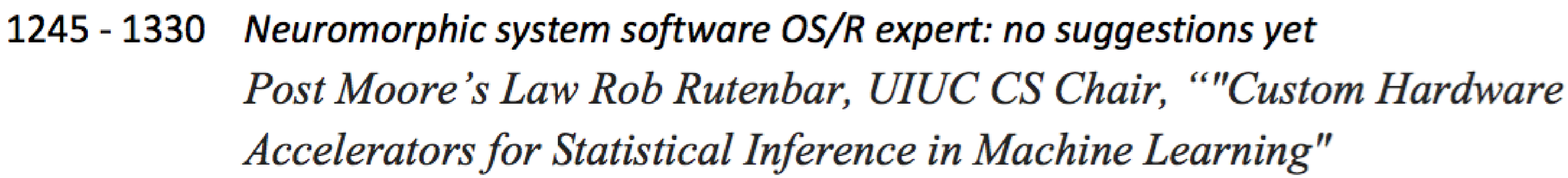
A Talk Across Two Domains….

Custom Accelerators

Machine Learning



Truth in Advertising Disclaimer..



1245 - 1330 *Neuromorphic system software OS/R expert: no suggestions yet*

Post Moore's Law Rob Rutenbar, UIUC CS Chair, ""Custom Hardware Accelerators for Statistical Inference in Machine Learning"

- **Today's talk**

Neuromorphic….?	**Perceptual AI apps**
System software + OS/R?	**Gates, flips flops, SRAM, wires…**
Post-Moore's Law	**End-of-Roadmap & Beyond**
Expert...?	**(TBD…)**

Accelerator Architecture: ***Why***?

- **Simple idea: Your app is too slow, or power hungry in software, so build *custom* hardware, *optimize* everything.**

Ex: Speech Recognition

voci

SUPPORT / CONTACT US

better data. better insights.

Use your speech and text analytics to gain insights about the voice of your customer, and surge ahead of the competition. Voci enables clients to have a 360° understanding of their customer to make knowledge-based decisions that are people and profit-friendly. Learn how Voci can help you get more out of your business intelligence.

BUSINESS INTELLIGENCE

what we do solutions partners about us

voices enlighten choices.

Every person, every voice, has a choice. Voci's mission is to enable enterprises to extract actionable insights from their voice data for powerful Speech Analytics, Business Intelligence, Electronic Discovery, and Compliance. Voci is the world leader in delivering automated speech transcription platforms of the highest accuracy, lowest latency, and best scalability. All for the lowest total cost of ownership. True business value for our clients.

automated voice transcription

speech analytics

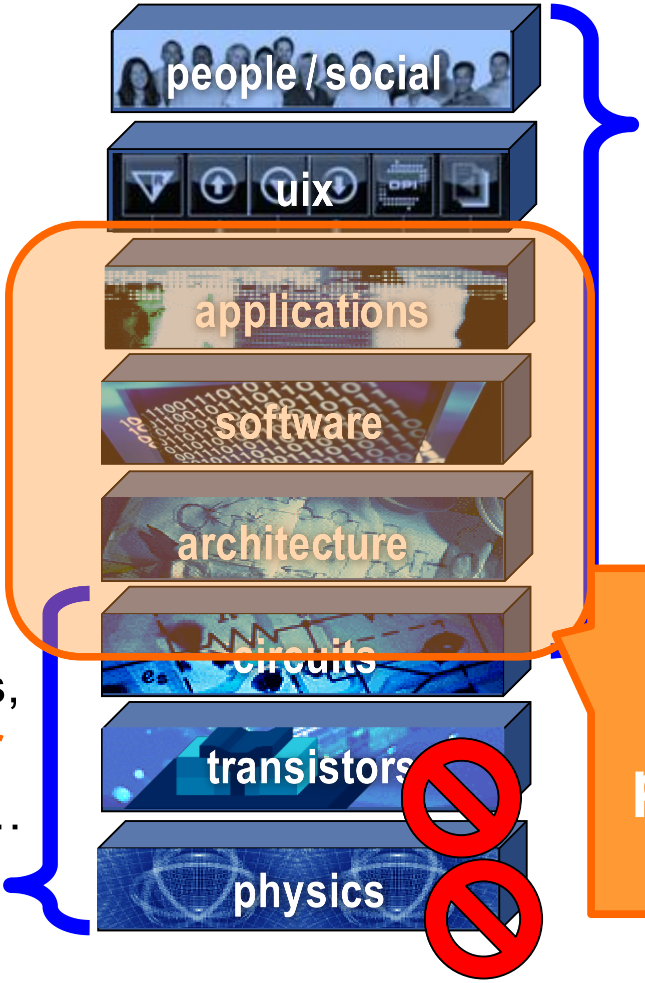
business intelligence

eDiscovery, compliance & risk

Accelerator Architectures: Why *Now*?

- **MOORE'S LAW**
  - 40+ yrs, every 2 years, transistors **2x smaller**
  - And **faster, cheaper**…
  - Law close to its **end**

The "stack"



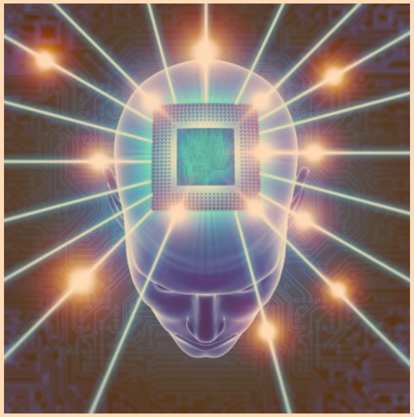
- **ATTN RISES**
  - Investment, interest moves **up the stack**
  - **Good for us in R&D**

My talk is this part of the stack

Hot Accelerator Areas

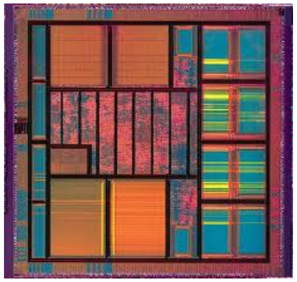
Bitcoin mining

High-freq trading



AI apps

- **Two broad “styles” of accelerators**



Fabbed

ASIC

Application Specific IC

Performance: Best

Cost: Worst

Configured

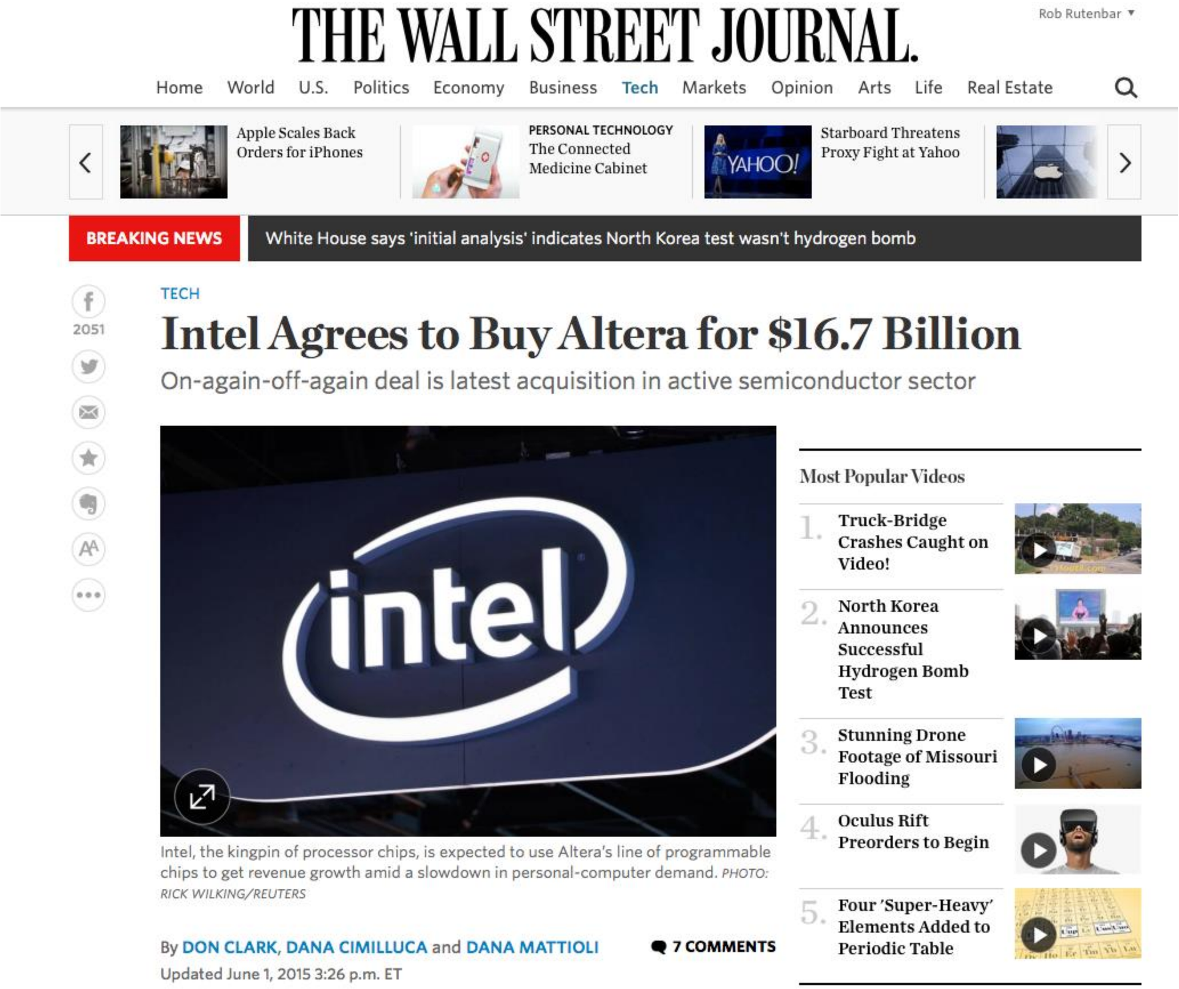
FPGA

Field Programmable Gate Array (i.e., ***reconfig****)*

Performance: Good

Cost: Cheap

Why FPGA Accelerators are Very Hot



THE WALL STREET JOURNAL.

Rob Rutenbar ▾

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

Apple Scales Back Orders for iPhones

PERSONAL TECHNOLOGY
The Connected Medicine Cabinet

Starboard Threatens Proxy Fight at Yahoo

BREAKING NEWS White House says 'initial analysis' indicates North Korea test wasn't hydrogen bomb

TECH

Intel Agrees to Buy Altera for $16.7 Billion

On-again-off-again deal is latest acquisition in active semiconductor sector

Intel, the kingpin of processor chips, is expected to use Altera's line of programmable chips to get revenue growth amid a slowdown in personal-computer demand. *PHOTO: RICK WILKING/REUTERS*

By DON CLARK, DANA CIMILLUCA and DANA MATTIOLI

7 COMMENTS

Updated June 1, 2015 3:26 p.m. ET

Most Popular Videos

1. Truck-Bridge Crashes Caught on Video!
2. North Korea Announces Successful Hydrogen Bomb Test
3. Stunning Drone Footage of Missouri Flooding
4. Oculus Rift Preorders to Begin
5. Four 'Super-Heavy' Elements Added to Periodic Table

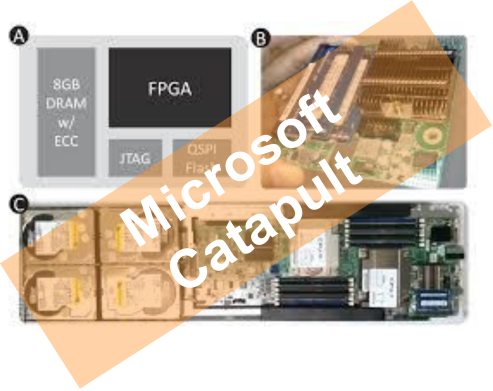
- **Moore's Law over…**
- **Intel needs "Plan B"**
- **2017: FPGA+CPUs all over data centers**
- **Upshot: Everybody can start building custom accelerators**

AI Opportunity Area: **Machine Learning** (ML)

- **Our apps can show us stuff, not understand stuff**
- **ML breakthroughs in recognition, classification**

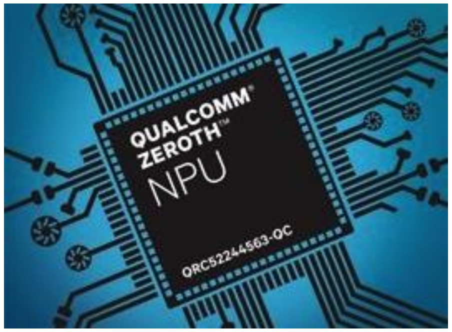
Activity Across Wide Spectrum

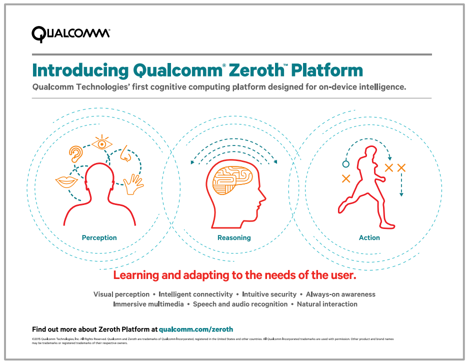
- **In Enterprise space**



(Next: Deep learning)

- **In Mobile space**




QUALCOMM

Introducing Qualcomm® Zeroth™ Platform

Qualcomm Technologies' first cognitive computing platform designed for on-device intelligence.

Perception
Reasoning
Action

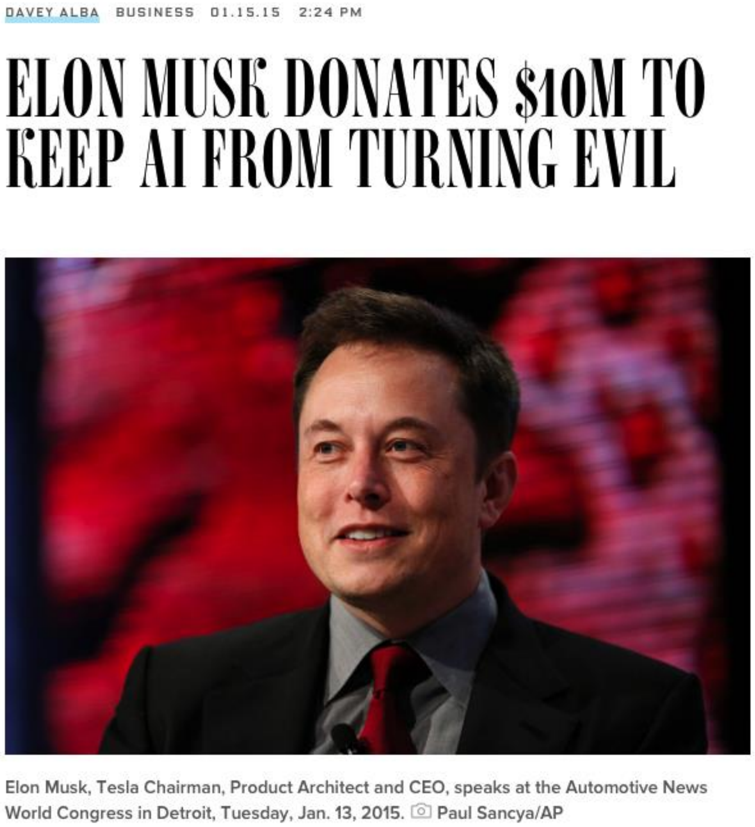
Learning and adapting to the needs of the user.

Visual perception • Intelligent connectivity • Intuitive security • Always-on awareness
Immersive multimedia • Speech and audio recognition • Natural interaction

Find out more about Zeroth Platform at qualcomm.com/zeroth

©2015 Qualcomm Technologies, Inc. All Rights Reserved. Qualcomm and Zeroth are trademarks of Qualcomm Incorporated, registered in the United States and other countries. All Qualcomm Incorporated trademarks are used with permission. Other product and brand names may be trademarks or registered trademarks of their respective owners.

- **In "hype" space**

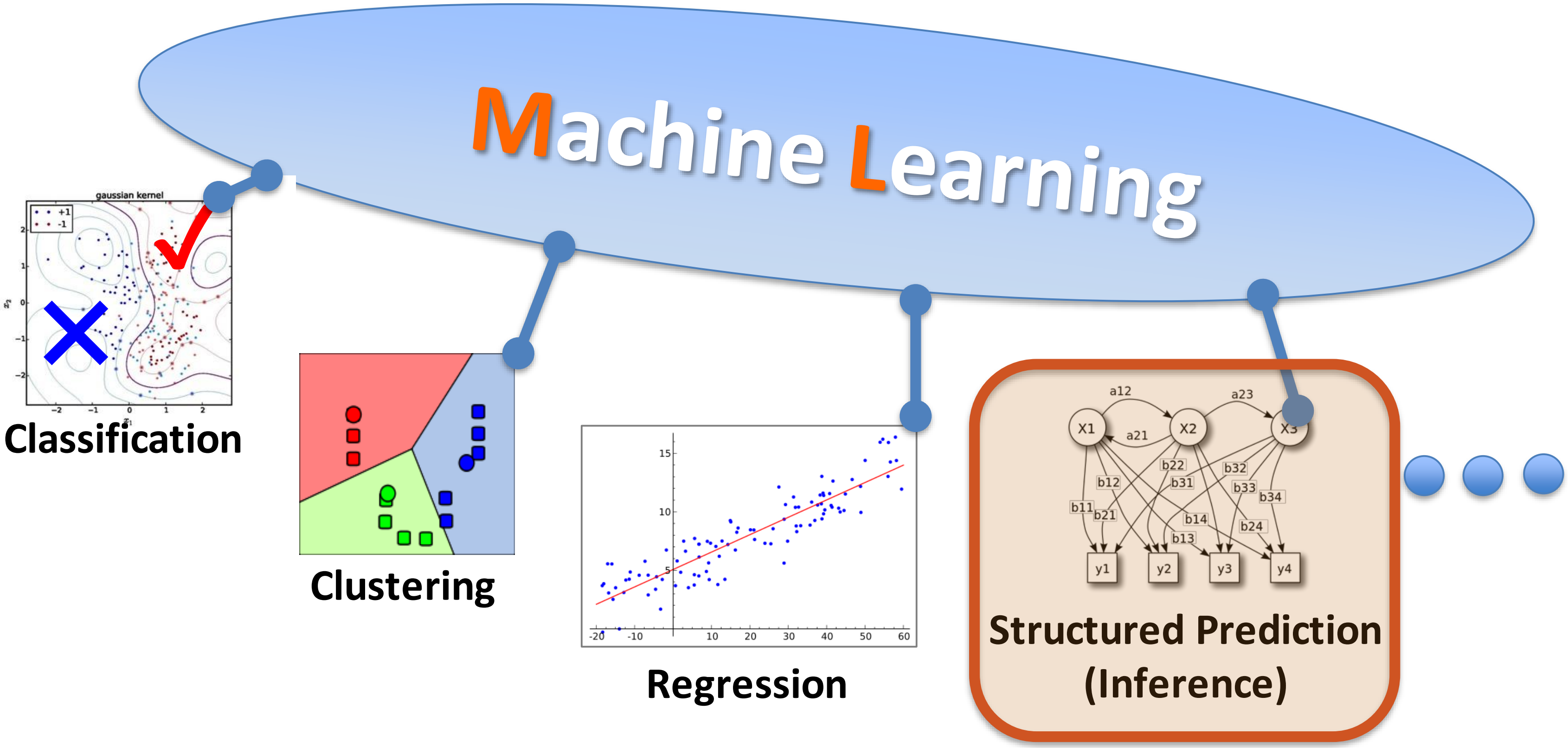

DAVEY ALBA BUSINESS 01.15.15 2:24 PM

ELON MUSK DONATES $10M TO KEEP AI FROM TURNING EVIL

Elon Musk, Tesla Chairman, Product Architect and CEO, speaks at the Automotive News World Congress in Detroit, Tuesday, Jan. 13, 2015. Paul Sancya/AP

Practical Reality: ML is *Huge* Area

- **Saying “We do ML” like saying “We do Math”…**
  - Critical to **focus** on “useful chunks” of ML domain



Machine Learning

Classification

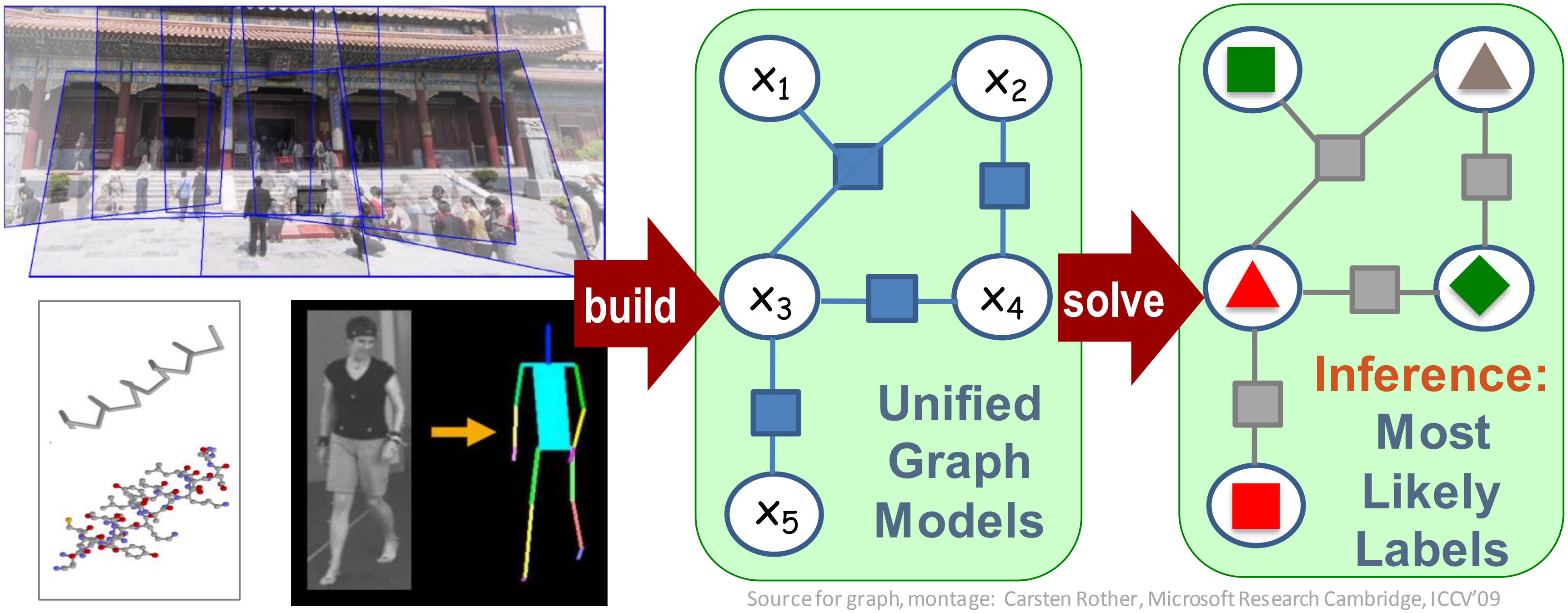
Clustering

Regression

Structured Prediction (Inference)

ML: Inference on Graphical Models

- **Important core ML technique, wide set of apps**
  - Nodes encode what we **observe/know**, how much we **believe it**
  - Edges encode **relationships** (joint dependencies/affinities)
  - Inference algorithms solve for **"most likely" labels** @ nodes

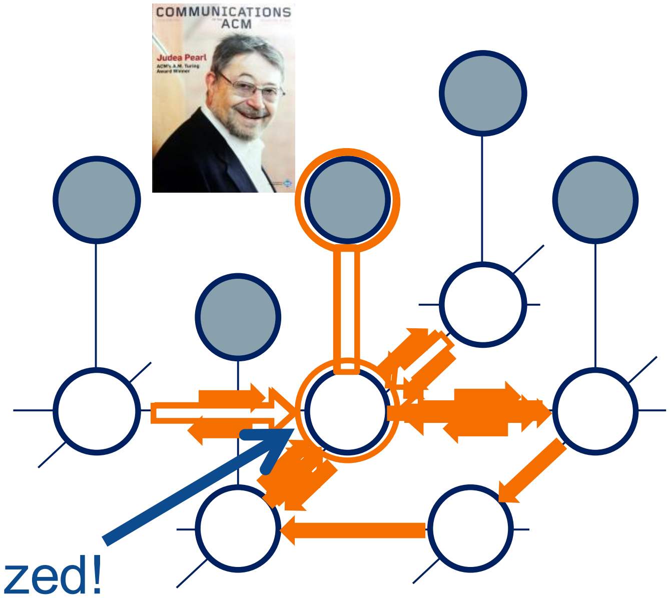


Apps

Source for graph, montage: Carsten Rother, Microsoft Research Cambridge, ICCV'09
Source for pose est: Pushmeet Kohil, Microsoft Research Cambridge, ibPRIA'11

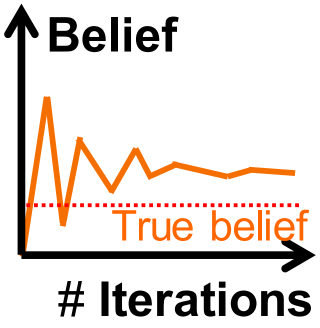
Inference **How**: Belief Propagation (BP)

- In BP, a node propagates **belief** to neighbors, iteratively via **messages**
  - **Message**: Based on what I know now, what do I tell to my neighbor?
  - **Belief**: What do I believe about labels based on my neighbors?”



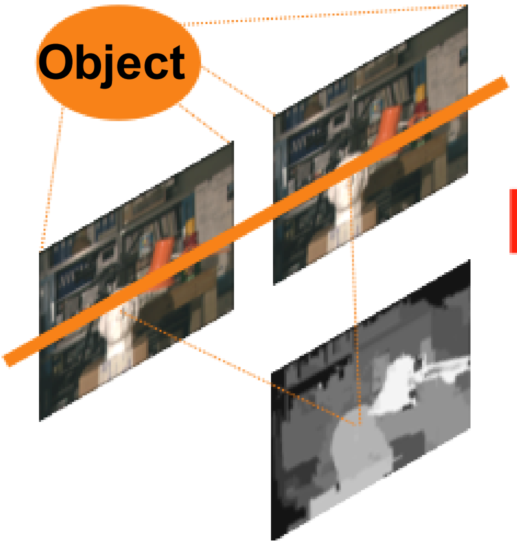
Over-emphasized!

- **Good**: BP on a tree **converges**
  - Most likely labels can be found after all inward/outward message passing is done

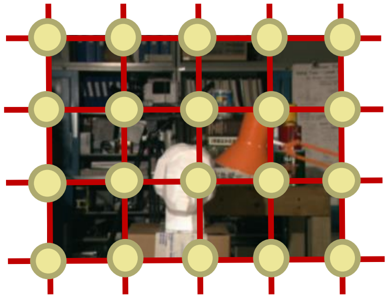


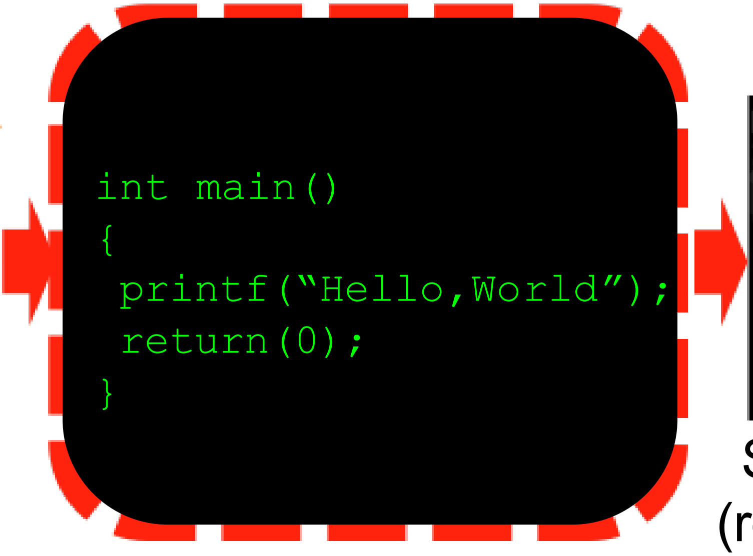
- **Bad**: BP on loopy graph **might not**

Stereo Matching as BP Inference



Formulate as **BP Inference**
on a probabilistic graphical model
one node per pixel





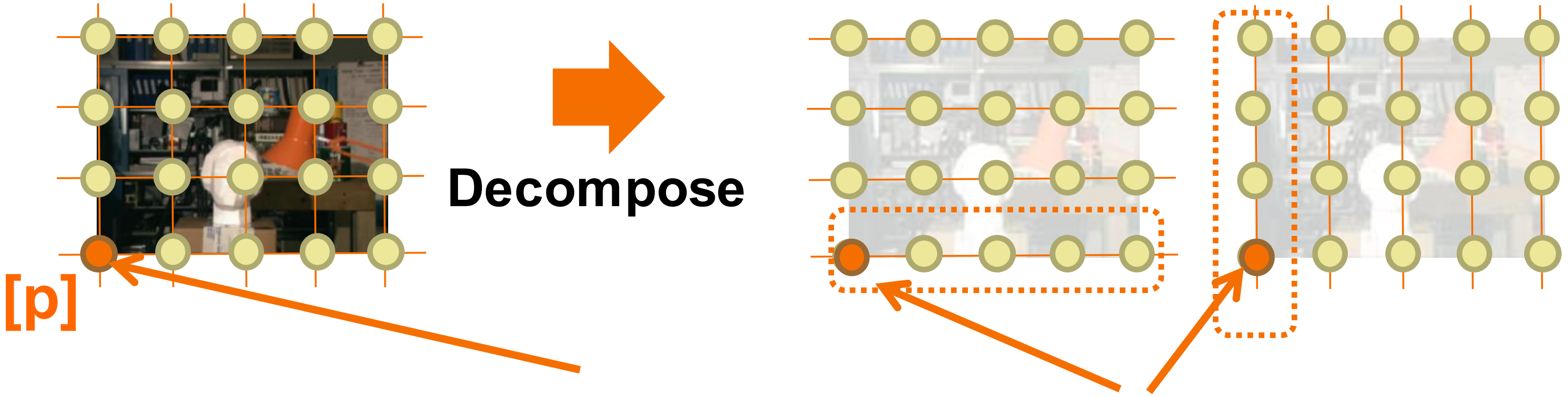
```
int main()
{
 printf(“Hello,World”);
 return(0);
}
```

Stereo disparity map
(reconstruct **3D depth**)

Design graphical model,
factors, and execute **BP**

Our BP: Sequential Tree-Reweighted (TRW-S)

- **Idea: Decompose a loopy graph to a set of trees, do inference sequentially across trees, recombine "right"**
  - [Kolmogorov PAMI'06]: Empirically v. good on loopy case; **slow**

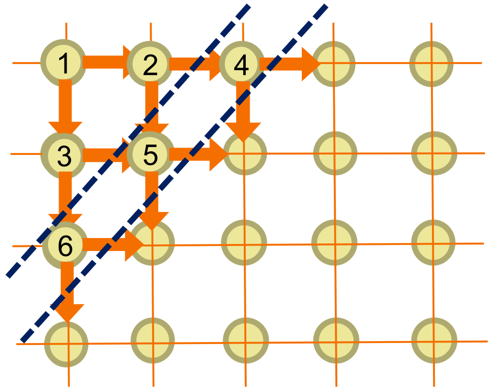


Recombine "smart": Outcome[p] = weighted sum from decomp

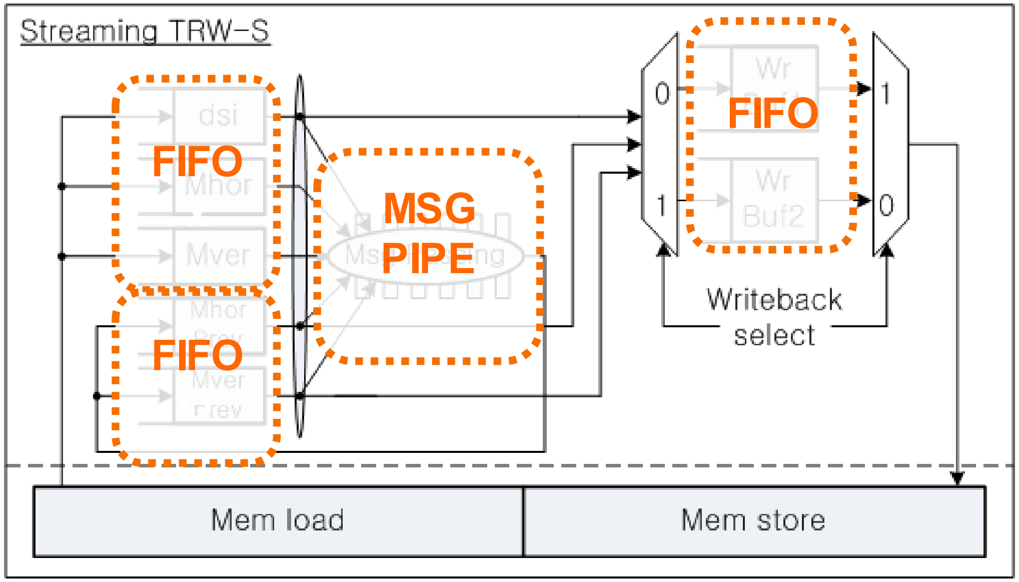
Problem: Sequential, across decomposed chains.
Bad (really bad) for hardware. Need a fix…

Fix: Streaming, ‘Diagonal Order’ Arch

Key: **Diagonal ordering of all message pass → parallelism**

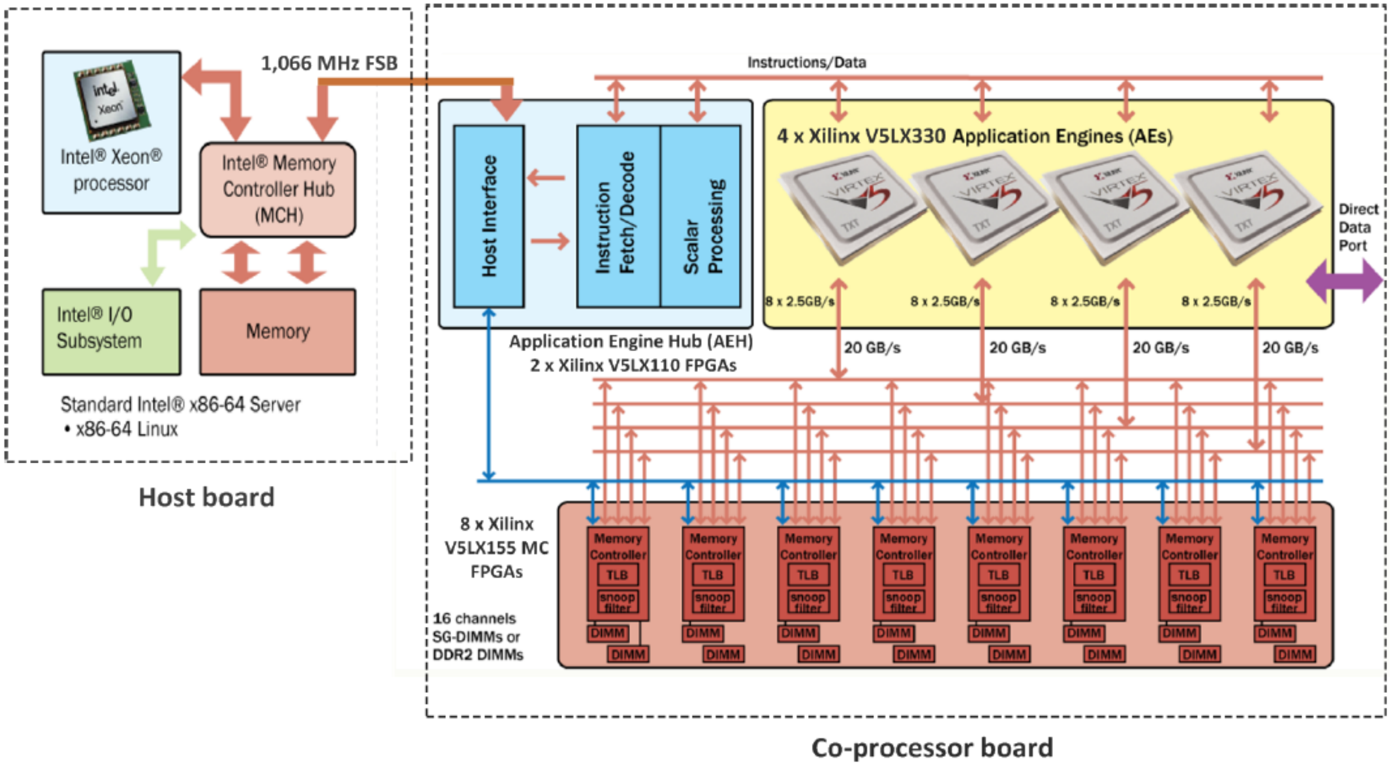


- **Decoupled, streaming arch**
- **Launch/retire 1 pixel/clock**
  - **Complete** label-set likelihood updates (~1Kb) for all labels
- **Deep pixel pipeline**
  - 14 stages deep
  - So: **14 pixels** “in flight” / clock



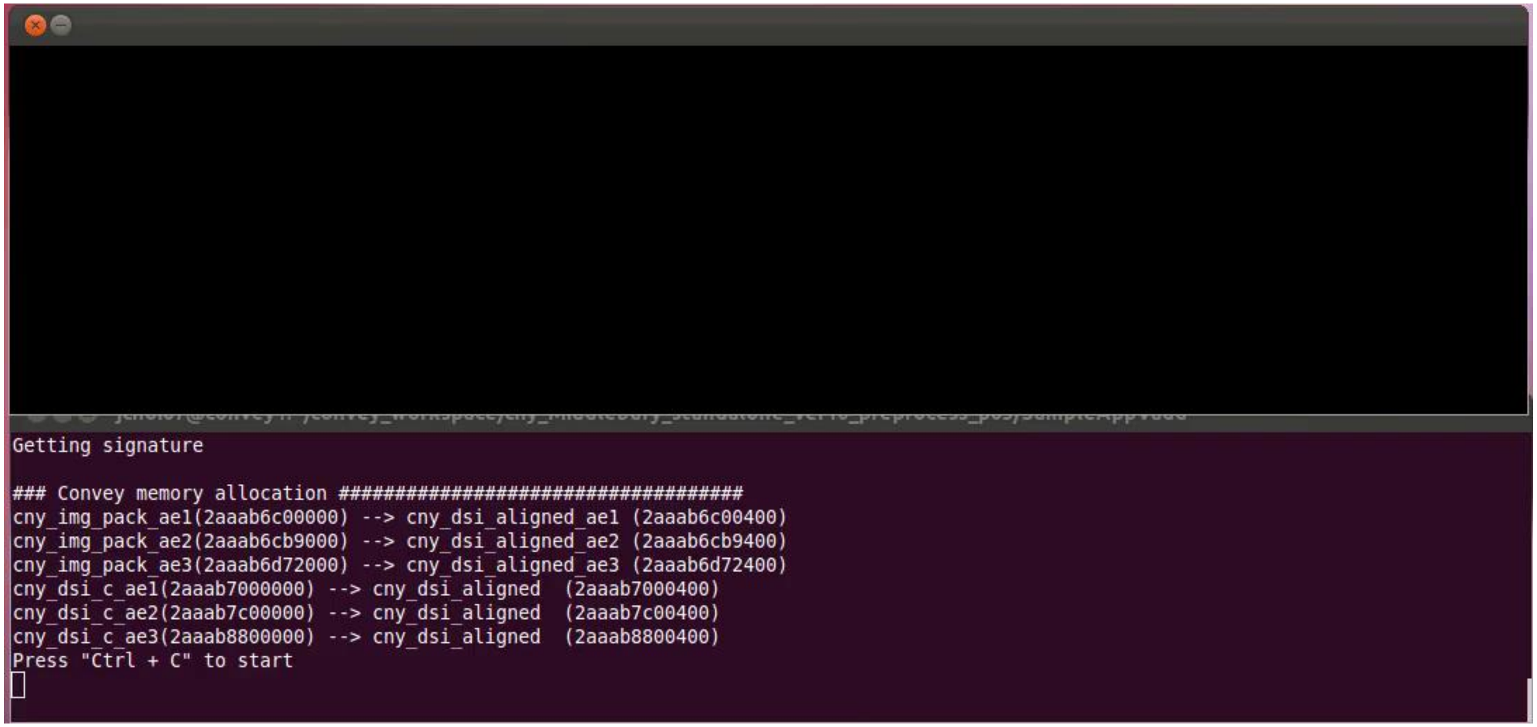
Our Platform: Hybrid CPU+FPGA

- **Our platform: Convey HC-1**
  - Intel Xeon + four Xilinx Virtex 5 (XV5LX330)
  - CPU-FPGA cache-coherent virtual memory system
  - Max memory BW: 1Kbit/cycle(~20GB/sec)/FPGA (runs @150MHz)



Stereo Performance Result in FPGA

- **Video frame rate BP inference [ISFPGA'13]**
  - Faster than competing software, GPU, ASIC published results



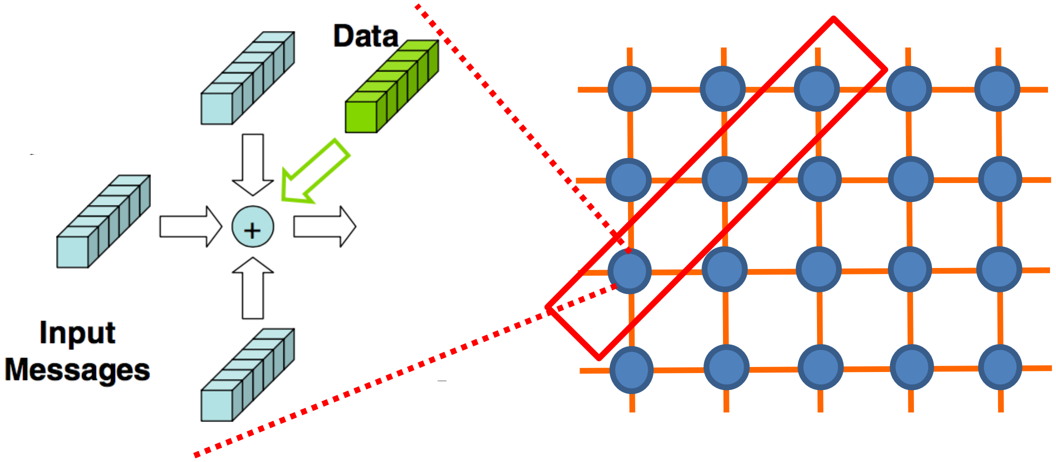
```
Getting signature

### Convey memory allocation ####################################
cny_img_pack_ae1(2aaab6c00000) --> cny_dsi_aligned_ae1 (2aaab6c00400)
cny_img_pack_ae2(2aaab6cb9000) --> cny_dsi_aligned_ae2 (2aaab6cb9400)
cny_img_pack_ae3(2aaab6d72000) --> cny_dsi_aligned_ae3 (2aaab6d72400)
cny_dsi_c_ae1(2aaab7000000) --> cny_dsi_aligned  (2aaab7000400)
cny_dsi_c_ae2(2aaab7c00000) --> cny_dsi_aligned  (2aaab7c00400)
cny_dsi_c_ae3(2aaab8800000) --> cny_dsi_aligned  (2aaab8800400)
Press "Ctrl + C" to start
```

(Details: 1 Xeon + 4 FPGAs; 20 TRW-S iter's/frame; QVGA 20fps; Scene-change-detect with message-reuse 'warmstart')

But, Need to Get Beyond “Hello World”

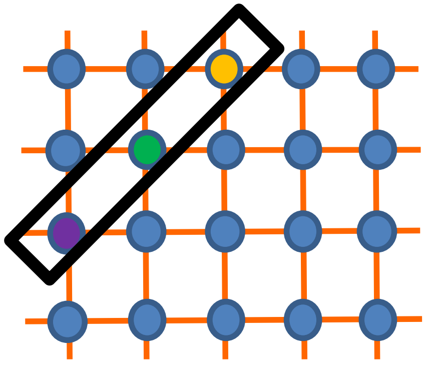
- **Valid criticism of point-accelerators: Narrow**
- **Problems with our current (Stereo) architecture:**
  - **Not configurable** to other inference problems
  - Pipelined, but **not scalable/parallel**
  - Uses **mem BW inefficiently** if |Labels| not multiple of 16



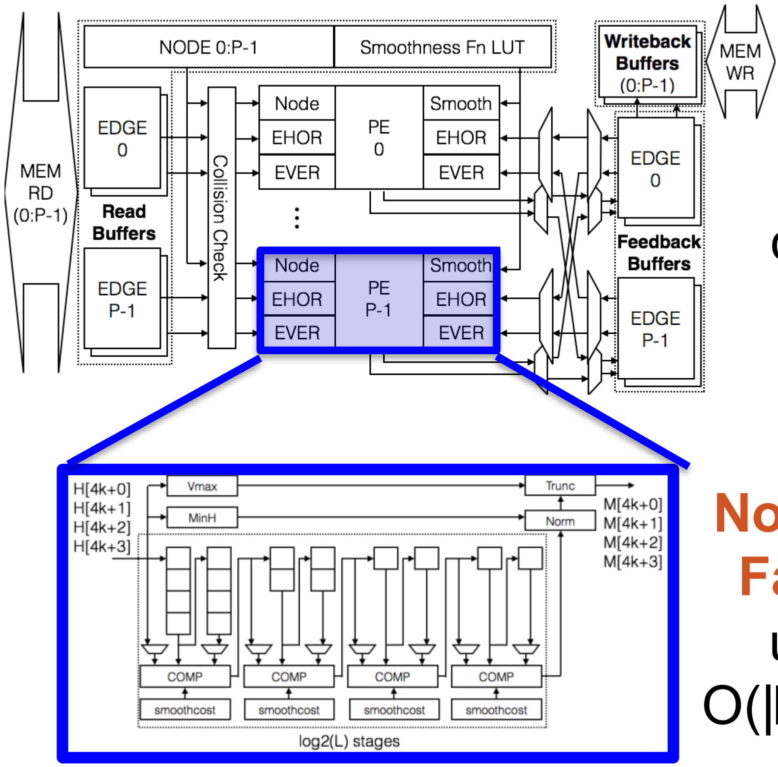
© 2016, R.A. Rutenbar

New: Scalable/Configurable BP Arch

- **Not just a pipeline any longer: *really* parallel…**



P Parallel processor elements (pixel streams)

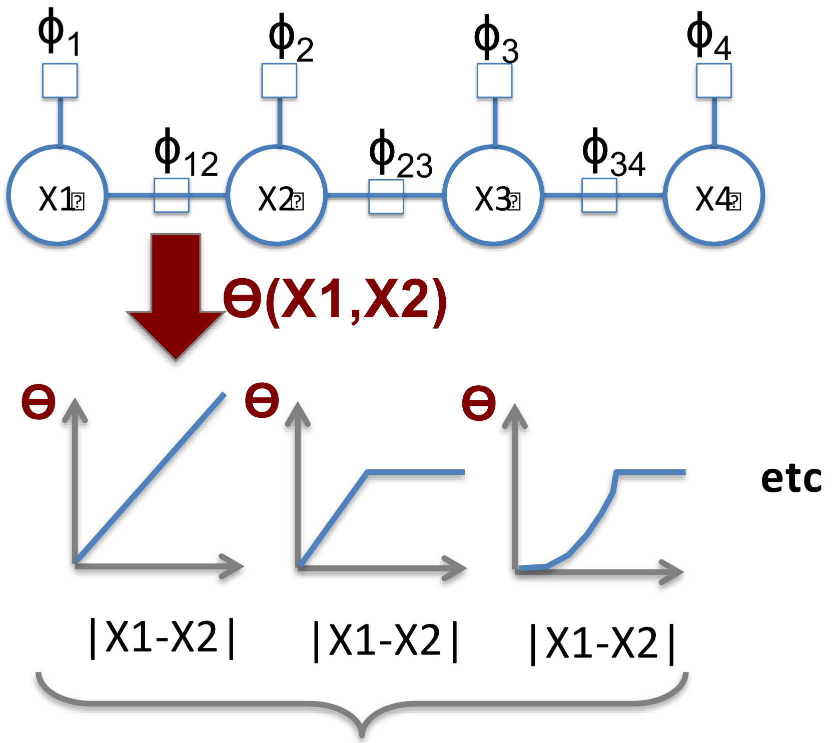


Efficient new memory subsystem overlaps BW and computation, checks for data conflicts

Novel, Configurable Factor-Evaluation unit removes the O(|labels|^2) complexity

© 2016, R.A. Rutenbar

Aside: What Does “Configurable” Mean?



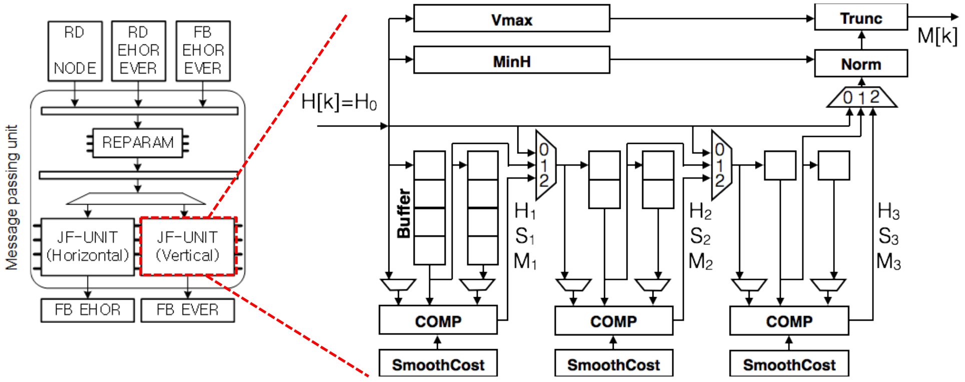
In vision, these pixel-to-pixel factors are called **Smoothness Costs,** reflect fact that pixels like to **agree**

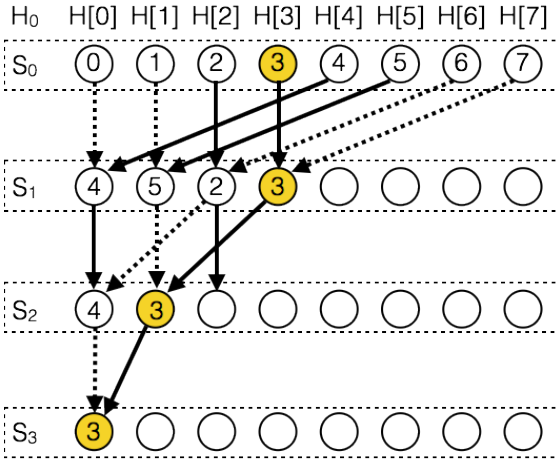
- **Essential fact**
  - **Standard forms** for “cost fn’s” for specific domains
  - We look at **computer vision**
- **We want to “hardwire” most common template for fn’s**
- **Unfortunate fact**
  - Might have **lots** of labels
  - MAP BP comp’s involve “MIN” op, **quadratic in |Labels|**

Fast Config Msg Passing: Jump Flooding

- **Problem:** BP msg computation **quadratic in L=|Labels|**
- **Solution:** Jump Flooding** BP msg **approx = L log(L)**
- **Analogy:** Like **"FFT"**, smart order for arith & comparisons

Configurable Jump-Flooding (JF-Unit) pipeline

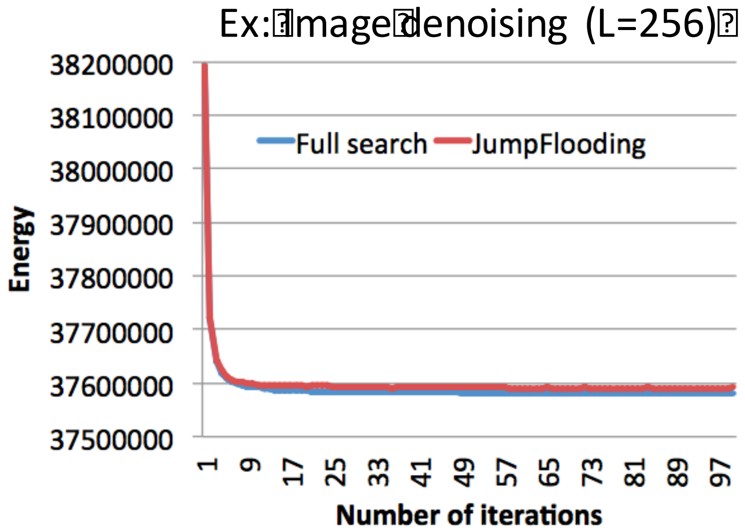




$S_3[0]=3$, $H_3[0]=H_2[1]=H_1[3]=H[3]$

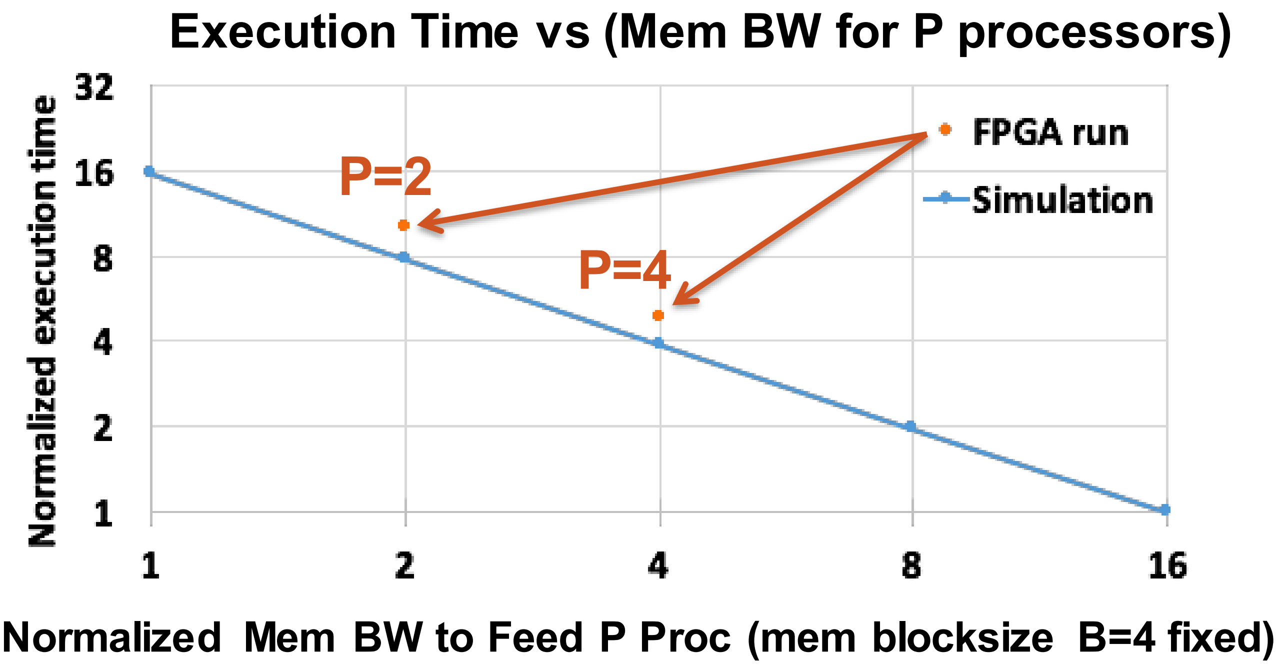
$M[0] = M_3[0] = H[3]+\theta[3,0]$

**[Rong, Tan, ACM Symp Int 3D, 2006]



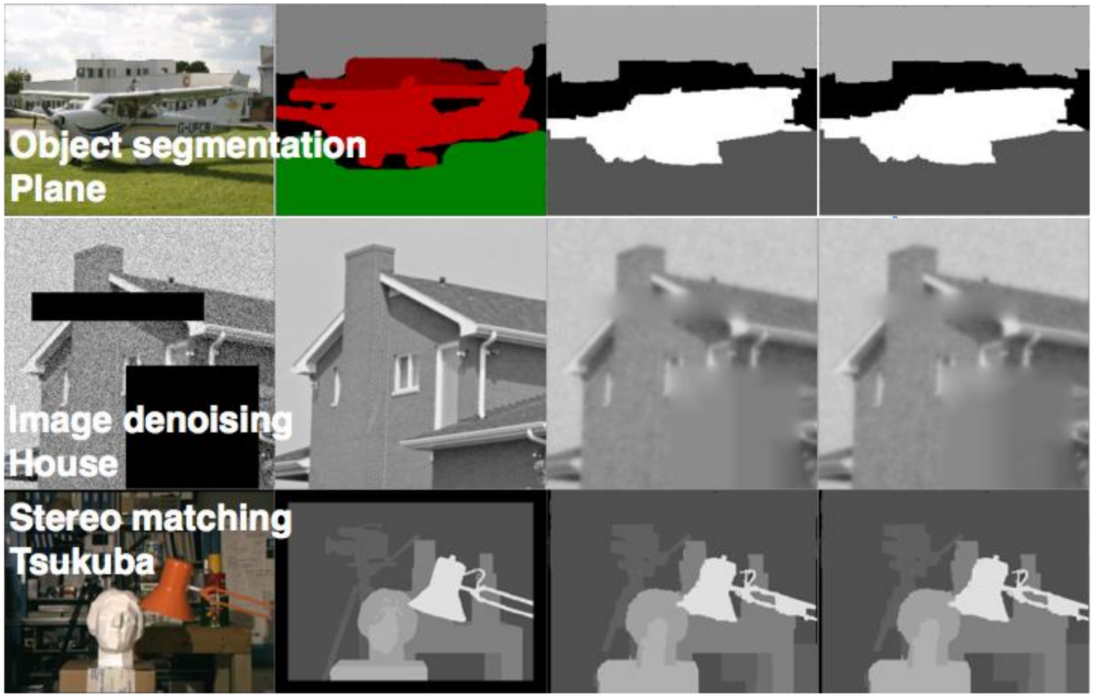
Positive Scalability Results

- **2, 4 PEs running (limited by Xilinx V5 size); sims 1-16 PEs**
  - Parameterized by “Bandwidth needed to feed P processors”
  - If we can *feed* the architecture – promising scalability



Results: Configurable BP Arch

- **12-40X faster than software (PE = 4)**
  - No loss of result quality
  - **1st "custom HW" to run >1 "standard**" ML inference benchmarks**



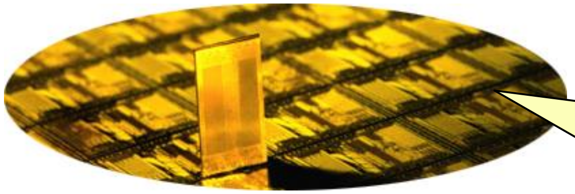
Input Ground Truth TRW-S SW Config Engine

**** Standard**
== *Middlebury*

© 2016, R.A. Rutenbar

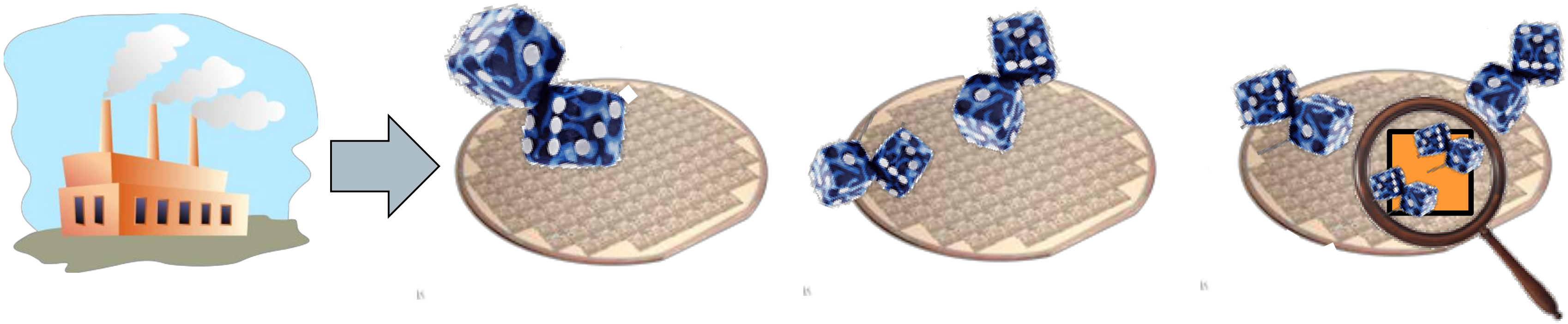
And Hardware BP Has *Yet More* Advantages

- **Dirty secret about "end-of-roadmap" & post-CMOS tech**



When basic switch is ~100 atoms wide
nothing is deterministic anymore

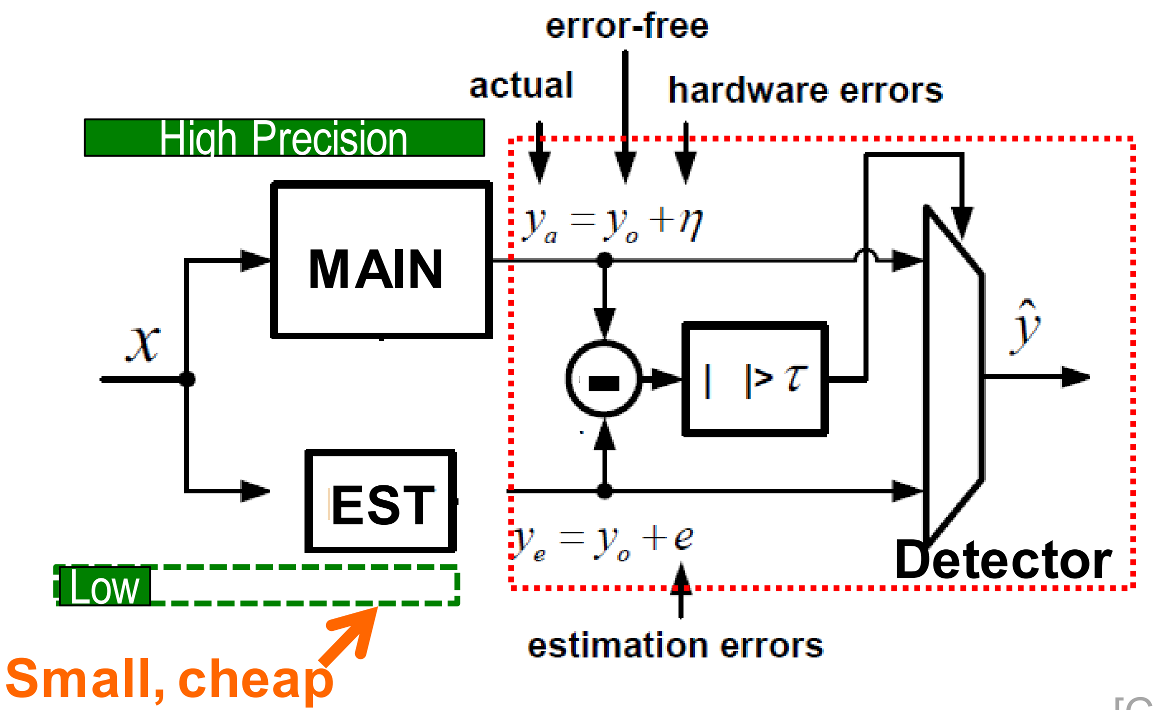
- **Every hardware behavior is a little smear of probability**



New problem: Resilience on a Stochastic fabric

BP Resilience via Algorithmic Noise Tolerance

- **BP's iterative character is already quite resilient…**
  - ...but not enough on really nasty stochastic fabrics
  - Studying **Algorithmic Noise Tolerance** (ANT) approaches



Small, cheap

if (MAIN ~ EST)
then MAIN
else EST

[Courtesy: N.Shanbhag UIUC]

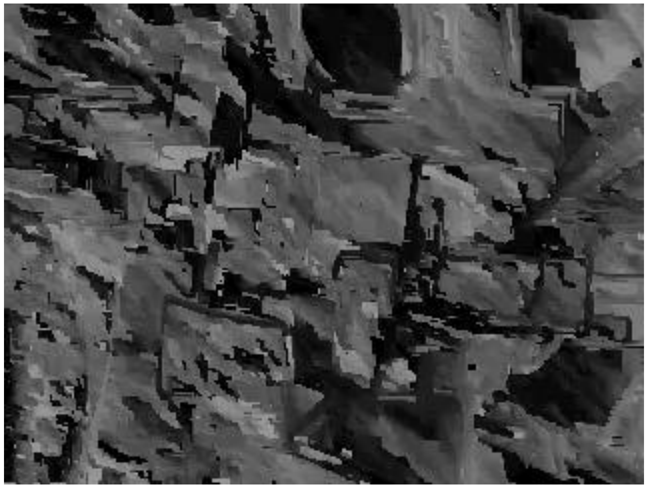
Big Result: BP Can Be Made Very **Resilient**

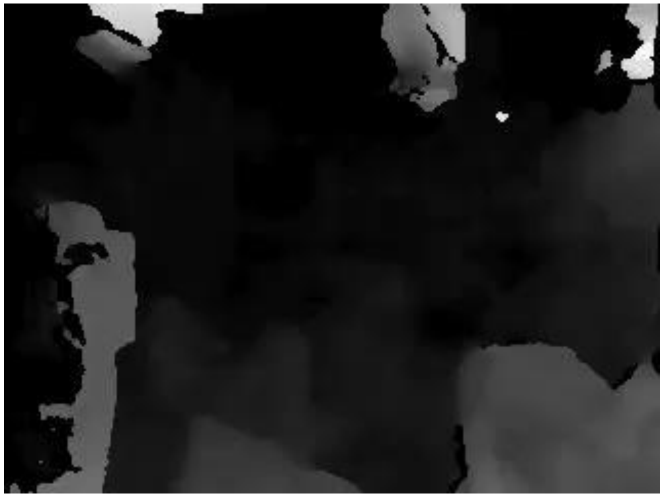
- **ANT "checker" HW && "pressure" from neighbor's beliefs**

LEFT

RIGHT

Synth Errors Added



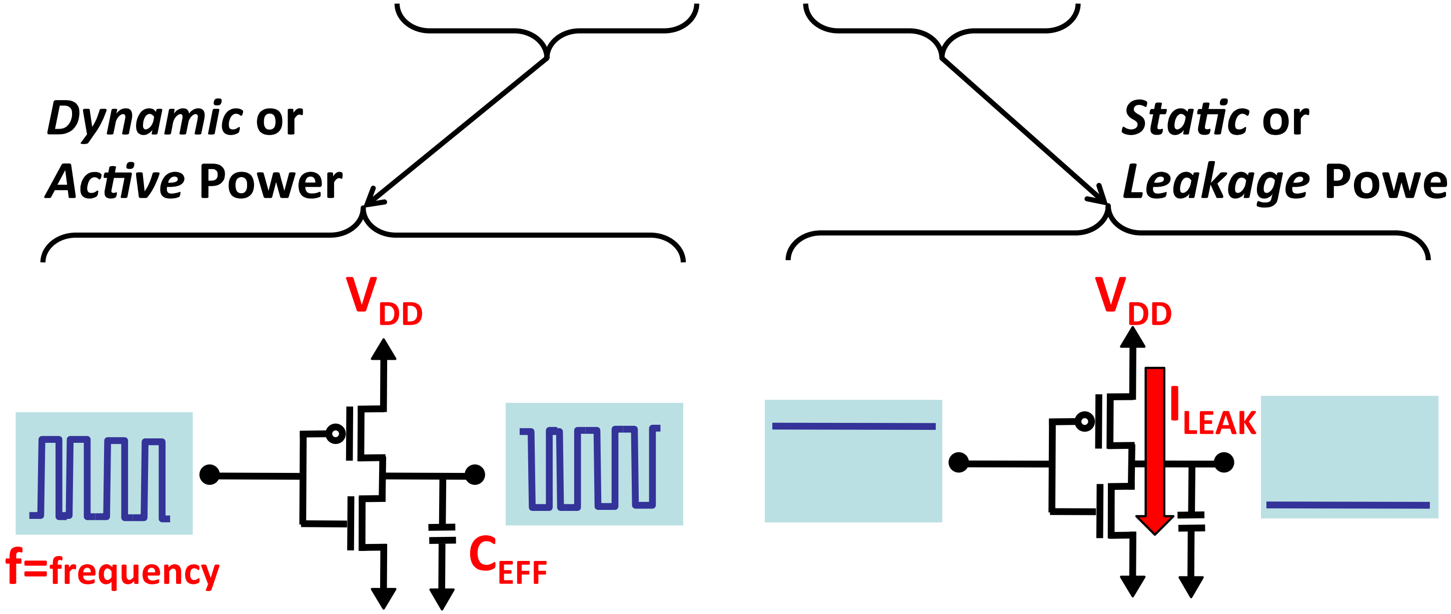


Smart Resilient Hardware

Even Better: Resilient → Lower Power

- **Why? A reminder from EE side: Iron Law of Power**

$$\textbf{Power} = \underbrace{C_{EFF}\, V_{DD}^{2}\, f}_{} + \underbrace{I_{LEAK}\, V_{DD}}_{}$$



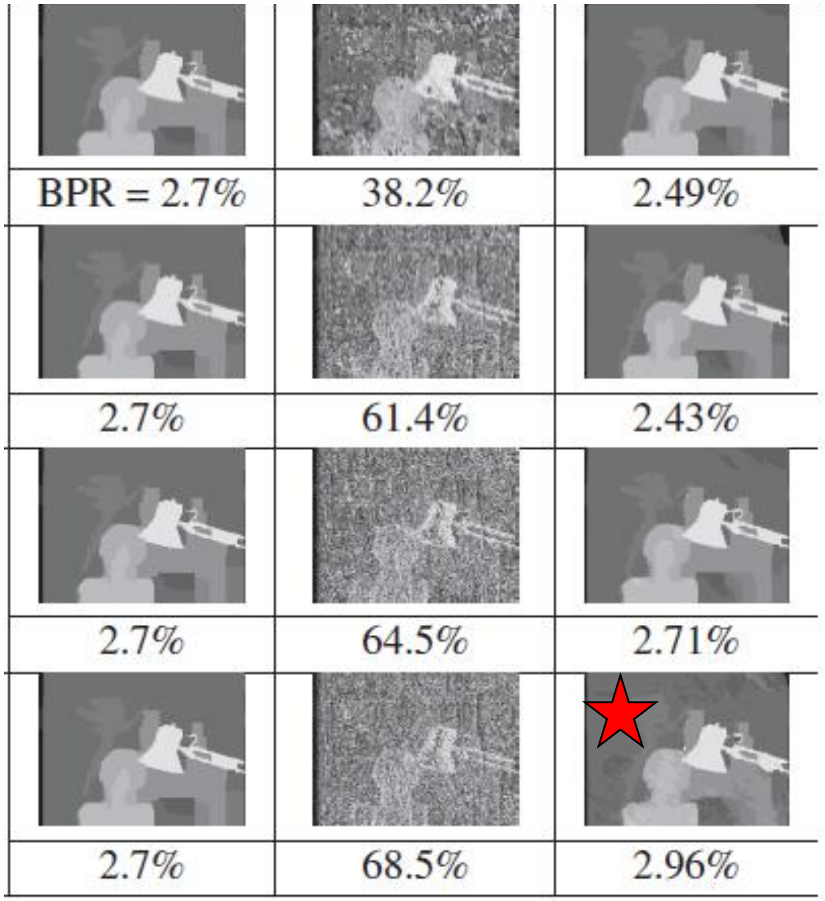
Dynamic **or *Active* Power**

Static **or *Leakage* Power**

Good News: To lower power, lower voltage VDD
Bad News: If VDD too small, everything breaks

Resilient→ Lower Power

- ANT mechanisms that "fix stuff" also handle low VDD



	Error-Free	Dumb	Resilient
VDD=1V	BPR = 2.7%	38.2%	2.49%
0.9V	2.7%	61.4%	2.43%
0.84V	2.7%	64.5%	2.71%
0.75V	2.7%	68.5%	2.96%

BPR =Bad Pixel Ratio (big=bad)

Voltage Over Scaling

43% power win!

Tech = 45nm

freq = 270MHz

VDD(nom)= 1.2V

Summary

- **Doing ML in custom hardware:**
  - Academically challenging && industrially relevant
  - Nice example of “cross over” from far ends of the computing space
  - Especially relevant as FPGAs go “maintstream” in enterprise & HPC
  - **Big need to partner with OS/R experts to make these practical…**

- **Lots of interest in this line of work:** ***intelligent systems***

DARPA intel ARM QUALCOMM IBM Micron