FPGA acceleration of Markov Random Field TRW-S Inference for Stereo Matching

MEMOCODE 2013 Design Contest

Jungwook Choi and Rob A. Rutenbar
{jchoi67,rutenbar}@illinois.edu

Oct. 19 2013

Outline

- Sequential tree reweighted message passing (**TRW-S**) for better MRF inference
- FPGA implementation of **streaming arch.** for high throughput TRW-S inference
- Performance results

  (+ Video stereo matching DEMO)

MRF infererence for Stereo Matching

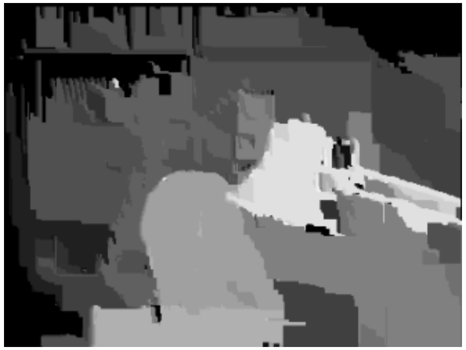
Maximum a posteriori (MAP)

Label assignments Observations

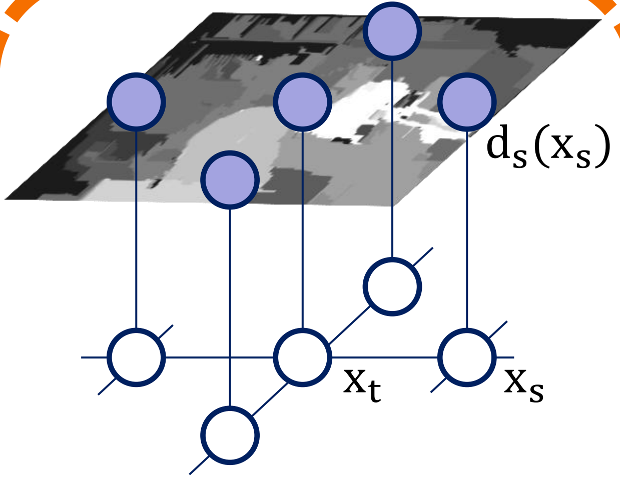
$$\operatorname*{argmax}_{\mathbf{x}} \underbrace{P(\mathbf{x}|\mathbf{y})}_{\text{Posterior}} = \operatorname*{argmax}_{\mathbf{x}} \underbrace{P(\mathbf{y}|\mathbf{x})}_{\text{Likelihood}} \underbrace{P(\mathbf{x})}_{\text{Prior}}$$

Energy minimization on Markov random fields (MRF)

$$\operatorname*{argmin}_{\mathbf{x}} \left(\text{Energy}\ (\mathbf{x}) \right)$$



3D depth map based on per-pixel likelihood of depth

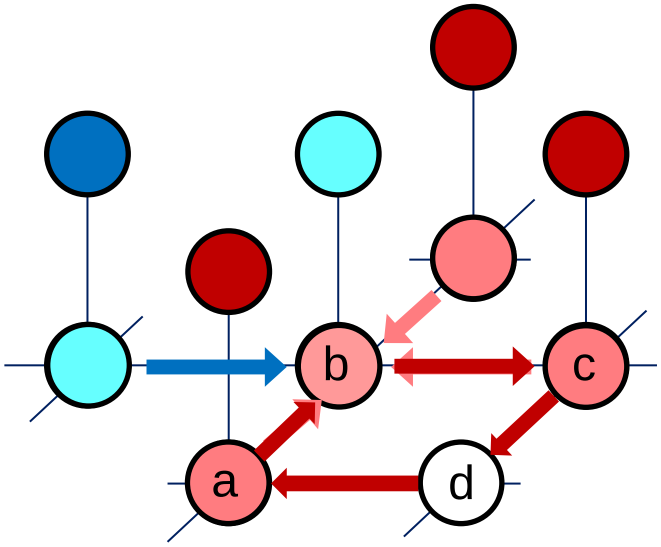


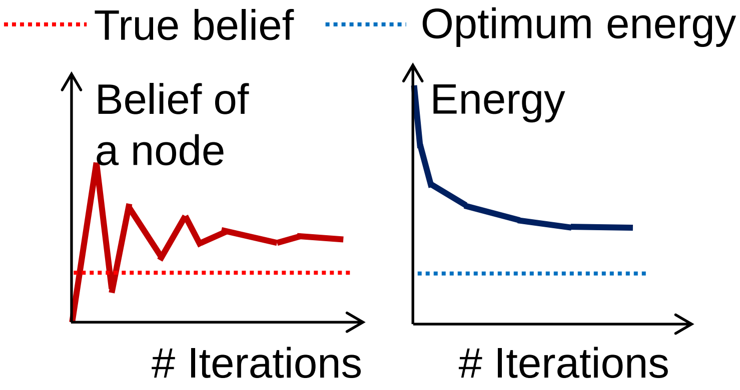
Likelihood → Data cost
Prior → Smoothness cost

3D depth map by MRF MAP inference

Loopy Belief Propagation

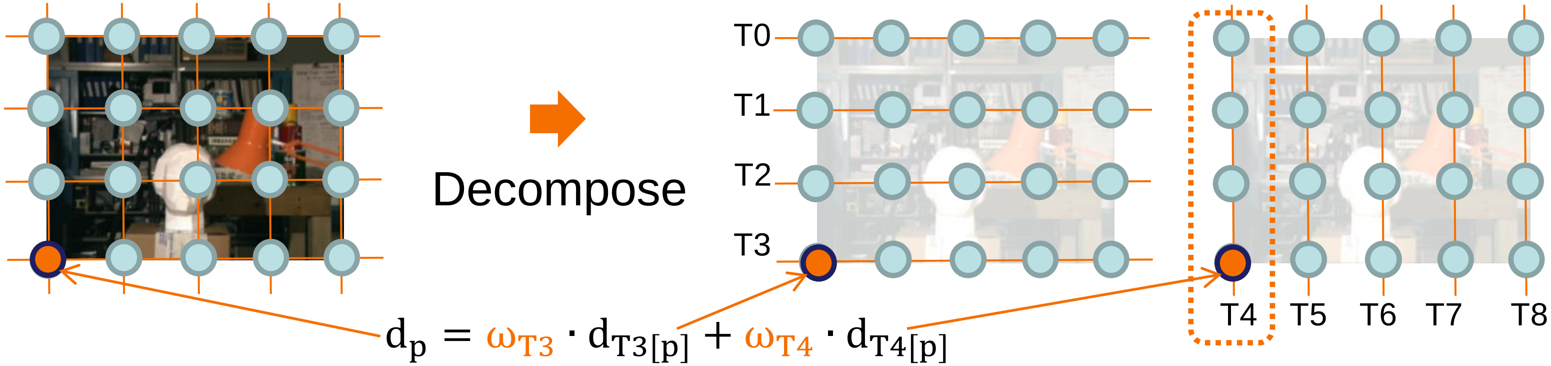
- "Based on what I hear, what label should I choose?"
  - Each node propagates "Belief" to neighbors via "message"
  - So called, "Message Passing" algorithm
- BP on **Tree**: Guaranteed to find optimal labeling
- BP on **Loopy graph**: No guarantee… (sometimes bad)





Better Alternative: Tree-Reweighted Message Passing

- Idea: decompose a loopy graph to a set of *trees*



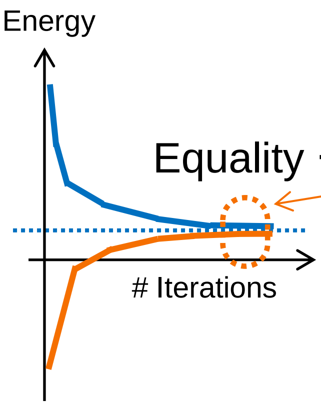
$$d_p = \omega_{T3} \cdot d_{T3[p]} + \omega_{T4} \cdot d_{T4[p]}$$

  - Energy is the weighted sum of tree energy

$$\text{Energy}(\mathbf{x}) = \sum_{\text{trees T}} \boxed{\omega_T} \cdot \boxed{\text{Energy}_T(\mathbf{x_T})}$$

Tree weight

Energy over a tree



Equality → optimum energy!

$$\min_{\mathbf{x}} \text{Energy}(\mathbf{x}) \boxed{\geq} \sum_{\text{trees T}} \omega_T \cdot \boxed{\left(\min_{\mathbf{x_T}} \text{Energy}_T(\mathbf{x_T})\right)}$$

Goal: minimize overall energy

New goal: maximize this lower bound

BP on a tree
➔ No loop!!

Sequential TRW (TRW-S)

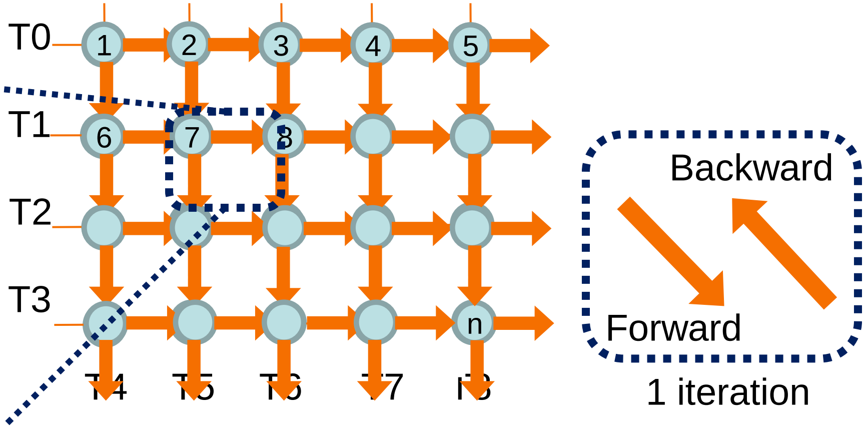
- New goal: max LB by ***two step message passing***
  - Averaging belief & message update
- Sequential message passing ➔ **Convergence property**
  - *Lower bound* is guaranteed not to decrease
  - ➔ More chance to find the optimum energy!!

- Averaging belief

$$\hat{\theta}_p(x_p) = \gamma_{pq} \cdot \left\{ d_p(x_p) + \sum_{s \in Nb(p)} M_{sp}(x_p) \right\}$$

- Message update

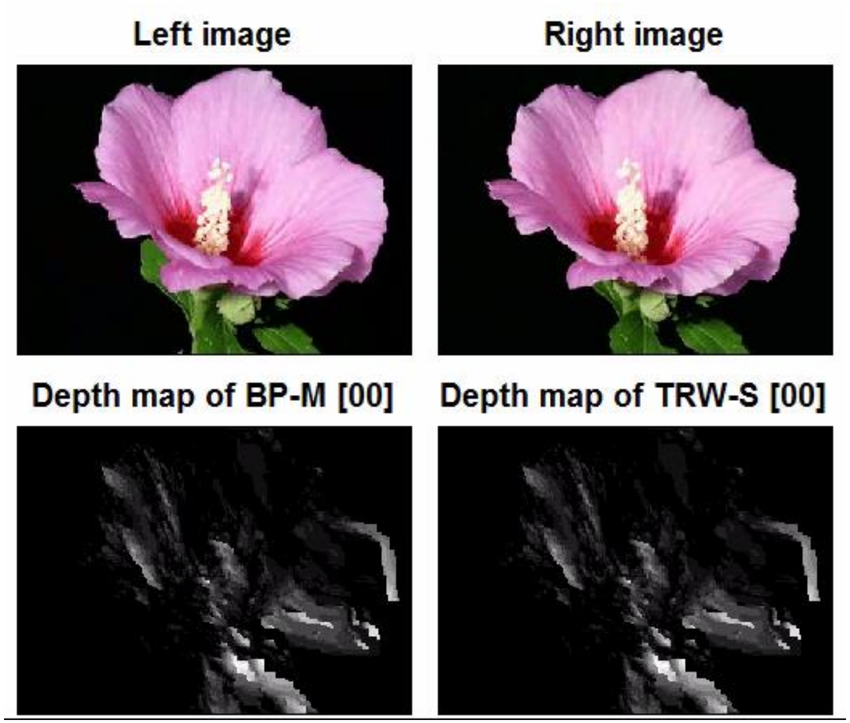
$$M_{pq}(x_q) = \min_{x_p} \{ (\hat{\theta}_p(x_p) - M_{qp}(x_p)) + V_{pq}(x_p, x_q) \}$$

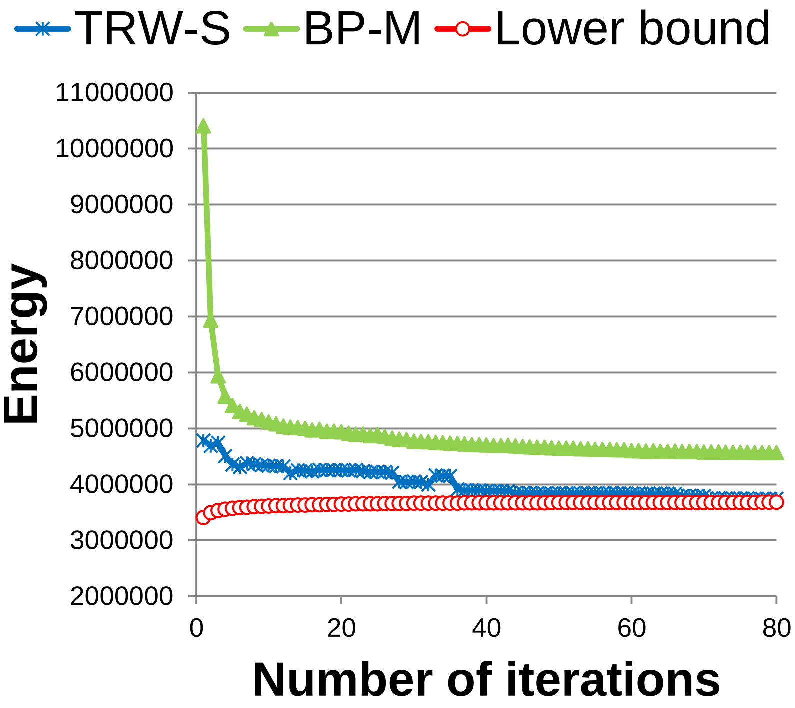


- Challenge : **parallelize** "sequential message passing"

Comparison: BP-M and TRW-S

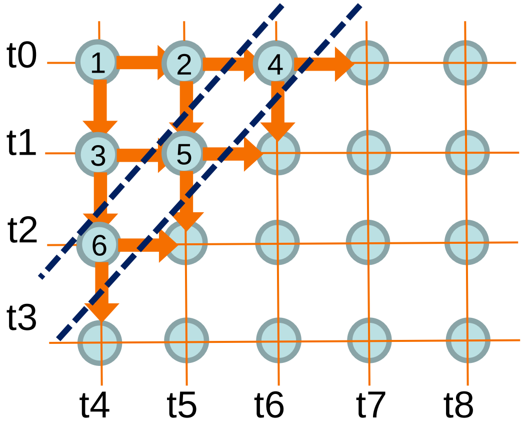
- Benchmark: Flower stereo images* (360x262x16 label)
  - Run message passing for 80 iterations

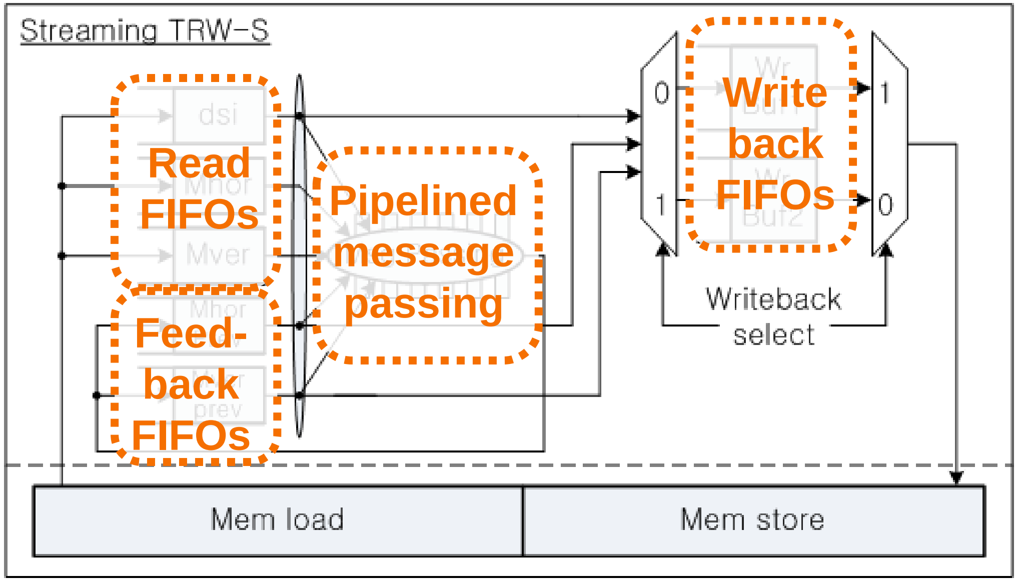




Streaming TRW-S HW Architecture

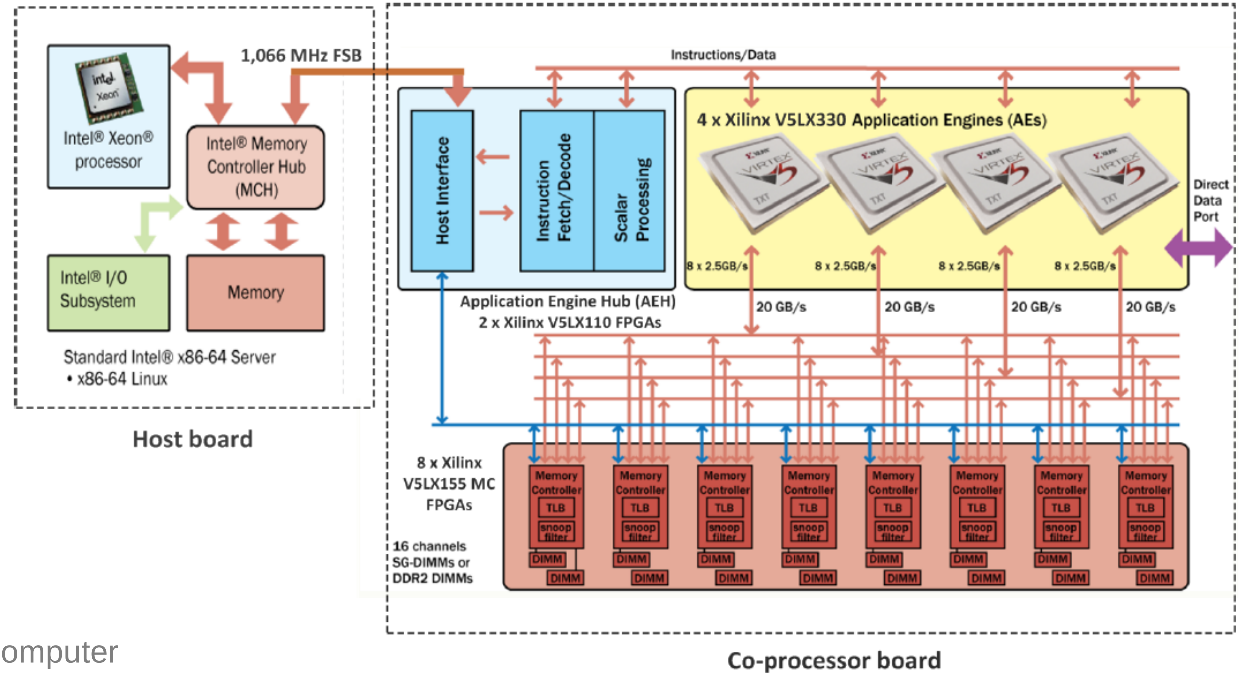
- Key: *diagonal ordering* of message passing for *parallelism*
- Decoupled, streaming arch.
- Launch/retire 1 pixel/clock
  - Complete label-set likelihood updates for all labels
- Deep pixel-proc pipeline
  - 14 stages deep
  - So: 14 pixels “in flight” / clock





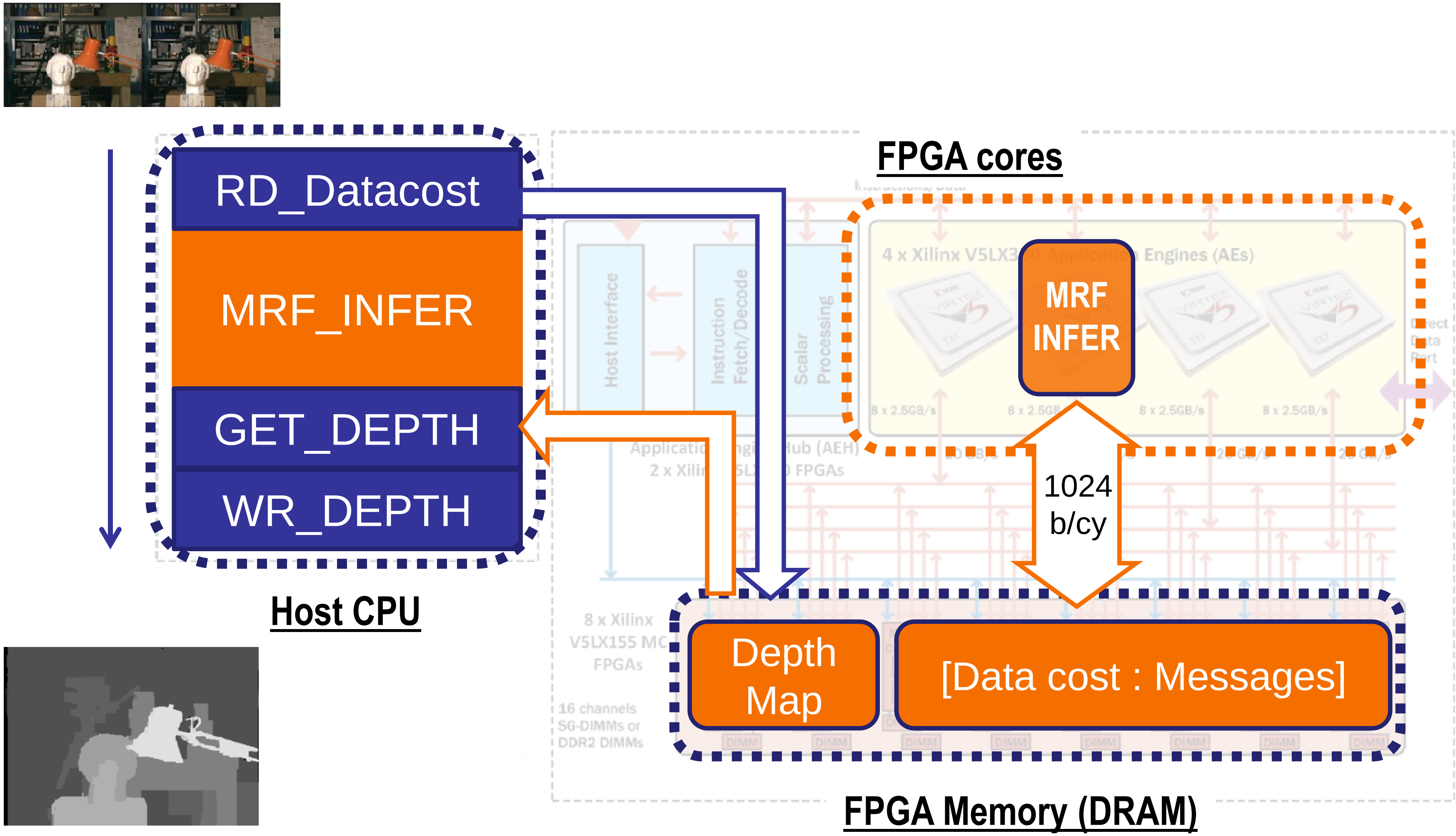
Hybrid CPU+FPGA Platform

- Our platform: Convey HC-1
  - CPU-FPGA cache-coherent virtual memory system
  - Max memory BW: 1Kbit/cycle(~20GB/sec)/FPGA (runs @150MHz)
  - Non-blocking FPGA function call



* Image from Convey computer

Stereo Matching Running on CPU+FPGA Platform



RD_Datacost

MRF_INFER

GET_DEPTH

WR_DEPTH

Host CPU

FPGA cores

MRF INFER

1024 b/cy

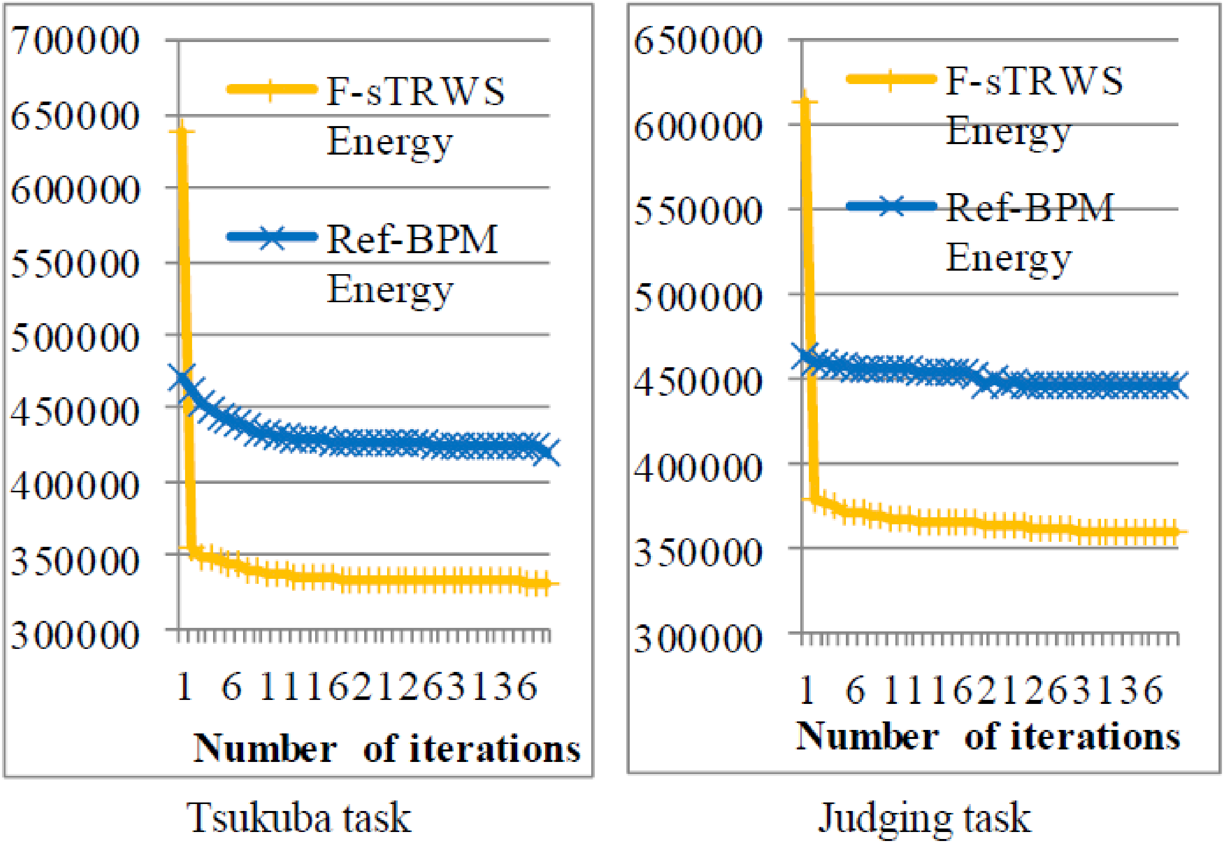
Depth Map

[Data cost : Messages]

FPGA Memory (DRAM)

Performance Results

- TRWS vs. BP: Min energy



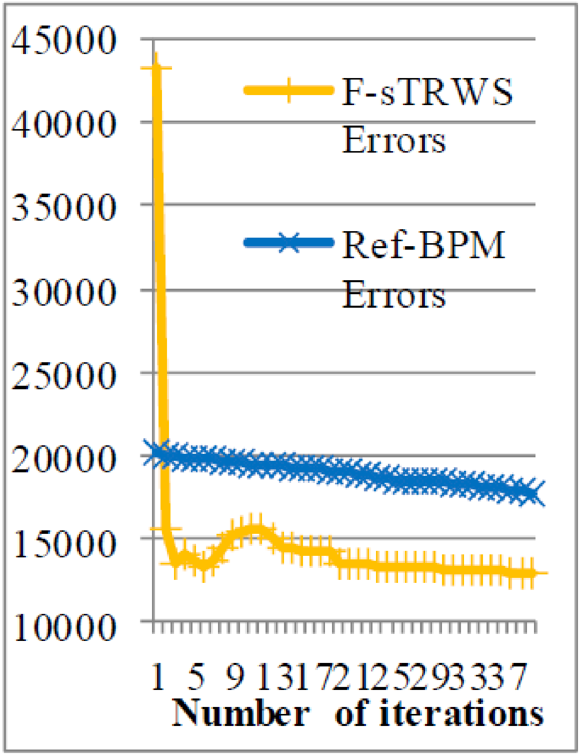
Tsukuba task

Judging task

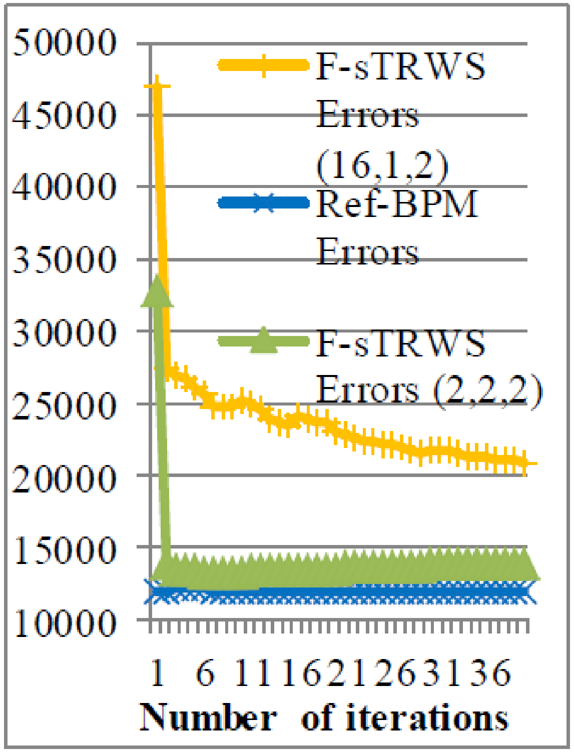
TRW-S achieves ***lower*** minimum energy in both cases

Performance Results

- TRWS vs. BP: Number of errors



Tsukuba task



Judging task

Tsukuba:
TRW-S takes two iterations to achieve lower #Error

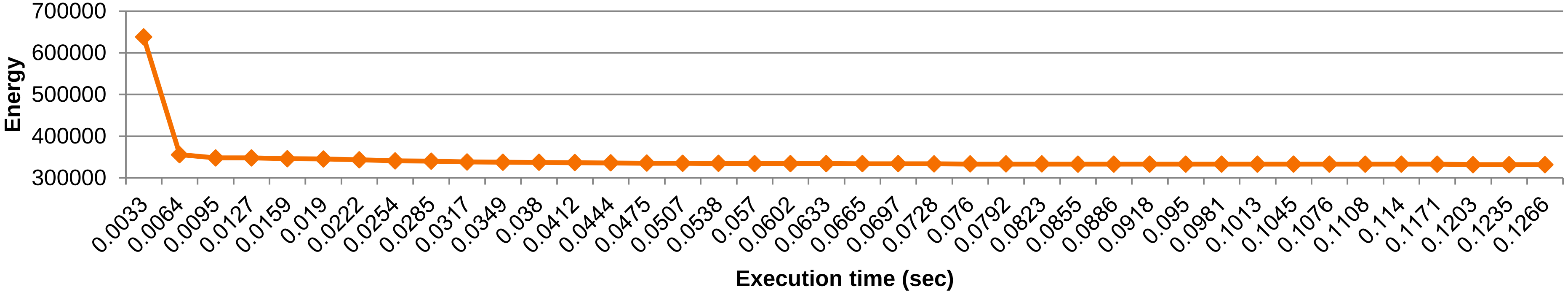
Judging:
TRW-S achieves comparable #Error, despite limit of MRF model

Performance Results

- Adjusted run time
  - Acc_BP: accuracy of reference BP, 40 iter.
  - Acc_TRWS: accuracy of TRW-S, 2 iter.

	Run time	Acc_TRWS	Acc_BP	Adjusted run time
Tsukuba	6.40 msec	82.1%	79.8%	6.40 msec
Judging	6.40 msec	84.3%	86.5%	7.43 msec

Performance Results



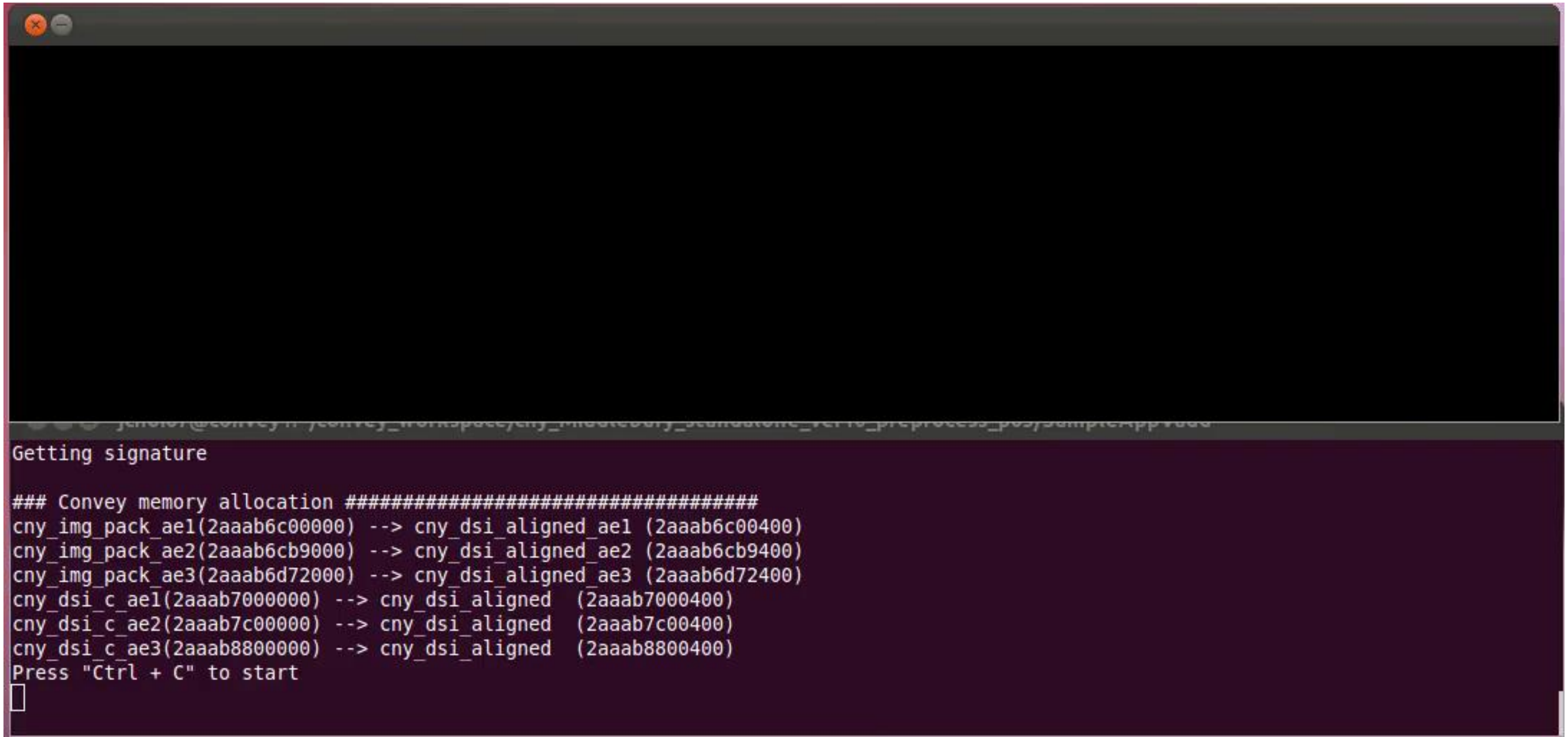
TRWS energy minimization over time (Tsukuba)

- It takes ~3.2msec to process 1 iteration
- Memory bandwidth utilization:
  - Data size: 64b x16(node data) x110,592(#nodes/iter) x2(fw/bw) x2(ld/st) = 56.6MB
  - Memory bandwidth: 1Kbit/cycle = 19.2 GB / sec
  - Memory bandwidth utilization: 56.6MB / (19.2GB/sec x 3.2msec) = **92.2%**

➔ Speed is bounded by memory bandwidth

Performance Result: Video Demo

- Function level pipelining + frame level parallelization
  - 3-frame parallel stereo matching, 80 iterations for each inference
  - Speed: Flower 12.3 frame/sec



```
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
```

Summary

- **Sequential tree reweighted message passing** (TRW-S) has been implemented in hybrid FPGA-CPU platform for high speed stereo matching
- TRW-S shows **faster convergence** to lower minimum energy than belief propagation (BP)
- **Pipelined streaming architecture** is exploited for high throughput message passing

	Run time	Acc_TRWS	Adjusted run time
Tsukuba	6.40 msec	82.1%	**6.40 msec**
Judging	6.40 msec	84.3%	**7.43 msec**

Thank You

Performance Result

Tsukuba (384x288,16)	**Real-time BP[*] [Yang 2006]**	**Tile-based BP[**] [Liang 2011]**	**Fast BP[***] [Xiang 2012]**	**This work**
GPU	NVIDIA GeForce 7900 GTX	NVIDIA GeForce 8800 GTS	NVIDIA GeForce GTX 260	N/A
# Iteration	(4 scales) = (5,5,10,2)	(B, T_I, T_O) = (12, 20, 5)	(3 scales) = (9,6,2)	$T_O = 5$
Time (msec)	80.8	97.3	61.4	**26.10**
				
Min. Energy	N/A	396,953	N/A	**393,434**

---

[*] Q. Yang, et al., “Real-time global stereo matching using hierarchical belief propagation,” *BMVC,* 2006.

[**] Liang, et al., “Hardware-Efficient Belief Propagation,” *IEEE Trans. Circ. Syst. Video Tech*, May 2011.

[***] X. Xiang, et al., “Real-time stereo matching based on fast belief propagation,” *MACH VISION APPL*, 2012