Asynchronous Stochastic Gradient Descent on GPU: Is It Really Better than CPU?

Yujing Ma, Martin Torres, Florin Rusu
University of California Merced

Problem

- Multi-core CPU or GPU, synchronous or asynchronous model updates, and data sparsity
- Hardware efficiency, statistical efficiency, and time to convergence
- Logistic Regression and Support Vector Machine

Parallel Gradient Descent

for $i = 1$ **to** #iterations **do**

1. Batch Gradient Decent

Compute gradient: ***(in parallel)***

for $t = 1$ **to** #tuples **do** $\vec{g} \leftarrow \sum_{\vec{x}_t, y_t} \overrightarrow{\nabla f}(\vec{w}, \vec{x}_t; y_t)$

Update model: $\vec{w}_i \leftarrow \vec{w_{i-1}} - \alpha\vec{g}$

2. Stochastic Gradient Decent

for $t = 1$ **to** #tuples **do** ***(in parallel)***

Compute gradient estimate: $\vec{g} \leftarrow \overrightarrow{\nabla f}(\vec{w}, \vec{x}_t; y_t)$

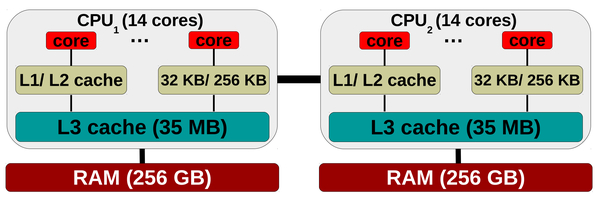
Update model: $\vec{w}_t \leftarrow \vec{w_{t-1}} - \alpha\vec{g}$

end for

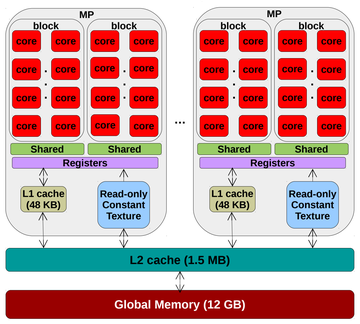
end for

Computing Architectures

- NUMA CPU Architecture (Intel Xeon E5-2660 v4)

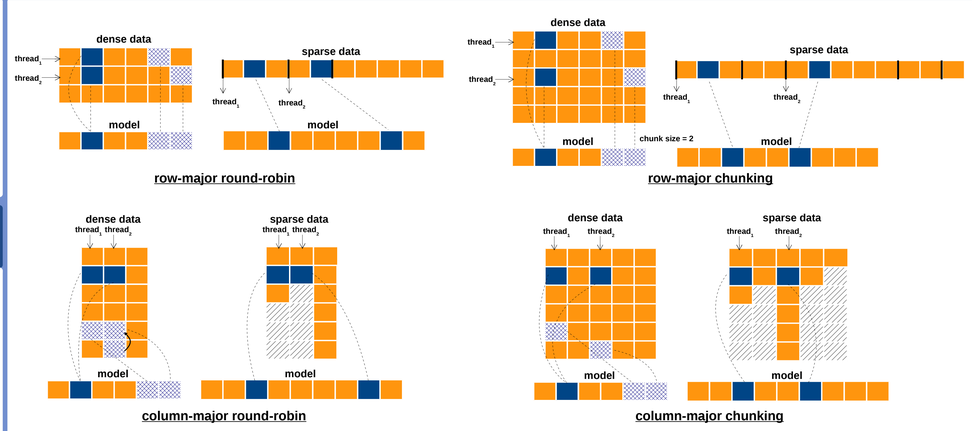


- GPU Architecture (NVIDIA Tesla K80)

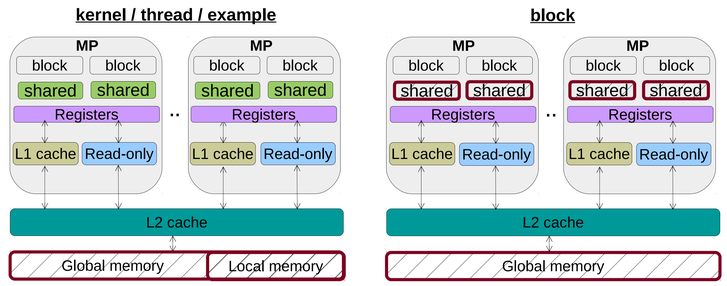


#MP	13
#cores/MP	192
L2 (MB)	1.5
Shrd/block (KB)	48
Warp size	32
#threads/MP	2048
#threads/block	1024

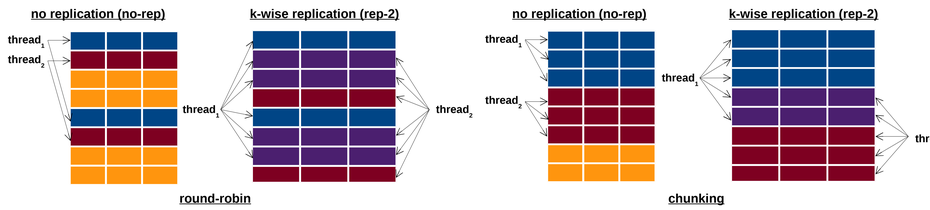
Data Access Path



Model Replication



Data Replication



Datasets

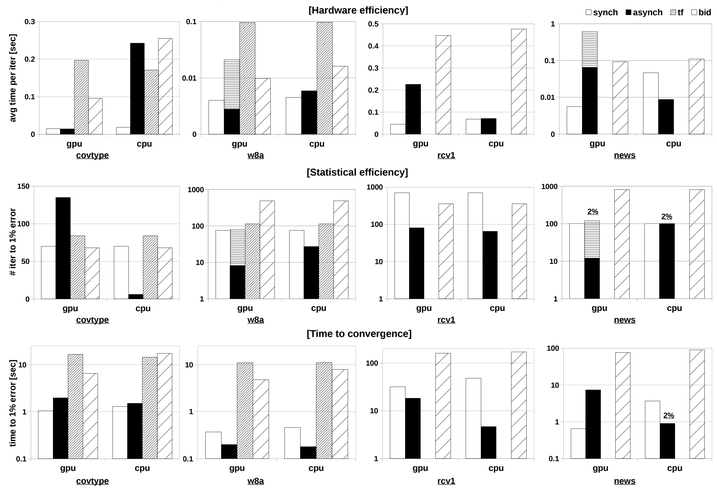
dataset name	#examples	#features	#NNZ/example (avg)	sparse size	dense size
covtype	581,012	54	54 (54)	485 MB	485 MB
w8a	64,700	300	0 to 114 (12)	4.4 MB	155 MB
rcv1	677,399	47,236	4 to 1,224 (73)	1.2 GB	256 GB
news	19,996	1,355,191	1 to 16,423 (455)	134 MB	217 GB

Experiments

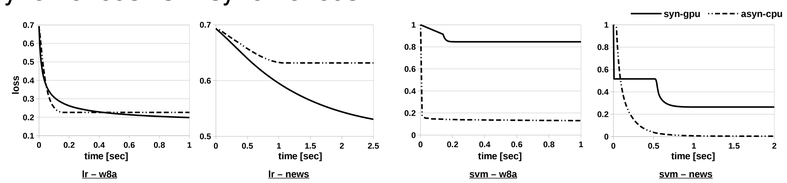
- GD performance to 1% convergence error (∞: not converge in 300 seconds)

algorithm (LR)	dataset	time-to-conv(sec) gpu	time-to-conv(sec) cpuS	time-to-conv(sec) cpuP	hardware(msec) gpu	hardware(msec) cpuS	hardware(msec) cpuP	statistical gpu	statistical cpuS	statistical cpuP
Synchronous SGD	covtype	1.05	145.11	1.29	15	2,073	18.42	70	70	70
	w8a	0.37	148.88	0.46	4.87	1,959	6.05	76	76	76
	rcv1	31.69	2,227	48.06	44.82	3,150	67.98	707	707	707
	news	0.65	240.21	3.68	6.37	2,355	36.08	102	102	102
Asynchronous SGD	covtype	1.97	0.60	1.51	15	150	251	135	4	6
	w8a	0.20	0.27	0.18	21 (2.8)	15	5.9	8 (80)	18	27
	rcv1	18.29	20.37	4.64	226	345	71	81	59	65
	news	7.35	5.47	∞	615 (65)	53	8.7	12 (∞)	103	∞

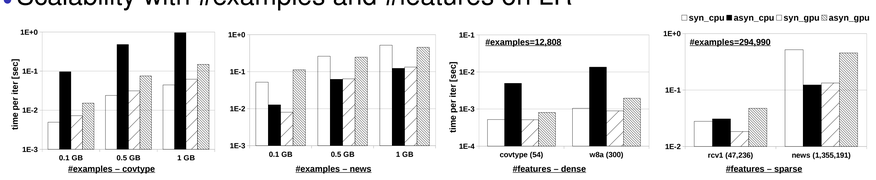
- CPU vs. GPU on LR (sync SGD, async SGD, TensorFlow and BIDMach)



- Synchronous vs. Asynchronous



- Scalability with #examples and #features on LR



- **GPU is the optimal architecture for synchronous SGD.**
- **CPU is optimal for asynchronous SGD.**
- **Choosing the better one is task- and dataset-dependent.**

{yma33, mtorres58, frusu}@ucmerced.edu