Adaptive Pipeline for Deduplication

Nankai-Baidu Joint Lab

Nankai University

Jingwei Ma, Bin Zhao, Gang Wang, Xiaoguang Liu

Nankai-Baidu Joint Laboratory

Parallel and Distributed Software Technology Lab

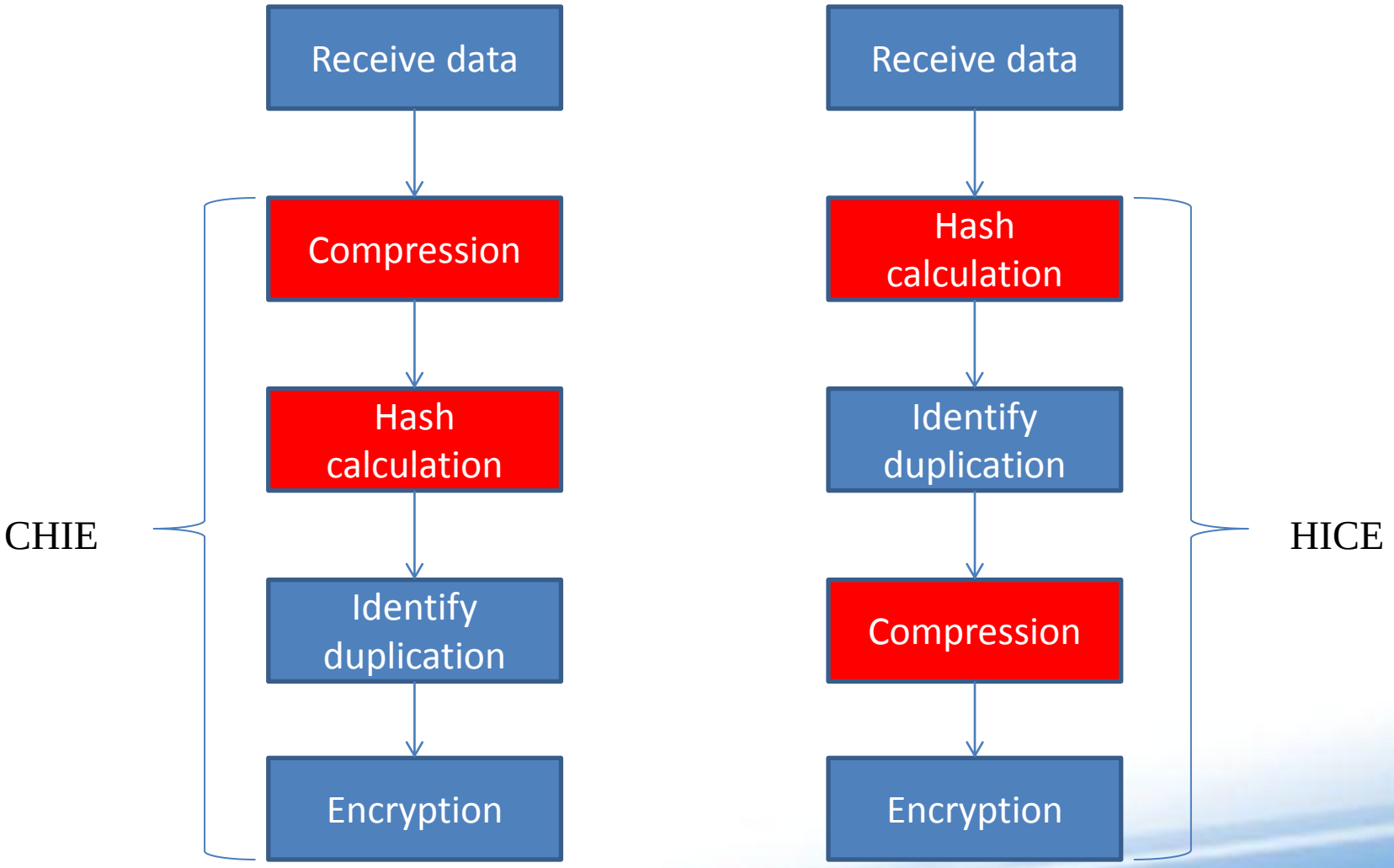
Deduplication Challenges

- I/O intensive
  - Fingerprint searches
    - Bloom Filter, Sparse Indexing, Extreme Binning
- Computation intensive
  - Fingerprinting, *Compression*, *Encryption*
    - GPU, SSE, Coprocessor

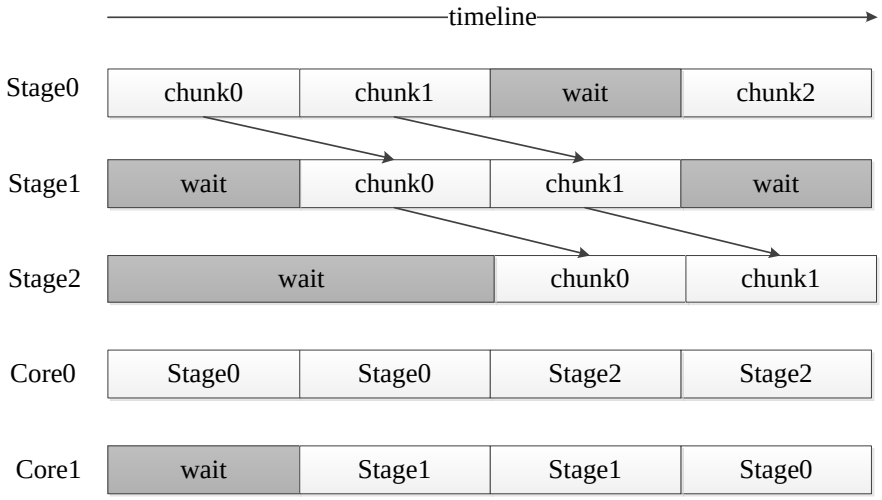
Our Motivation

Can the order of the sub-tasks in deduplication affects the throughput?

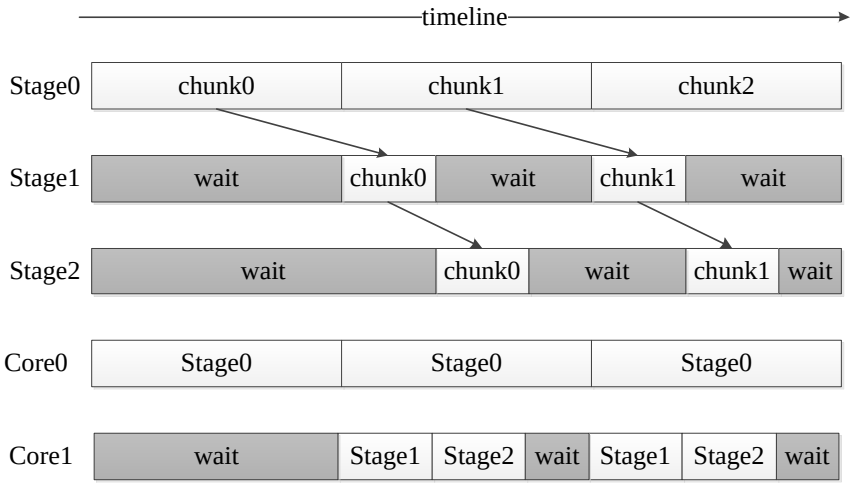
Two Orders of Sub-tasks



Two Situations



Balanced, The CPU is exhausted by the program



Unbalanced, The CPU is not exhausted by the program

Balanced Situation

The total time of which is shorter?

CHIE

$$\frac{S}{C} + \frac{S \times R_C}{H} + T_{Duplication\ identification} + T_{Encryption}$$

VS

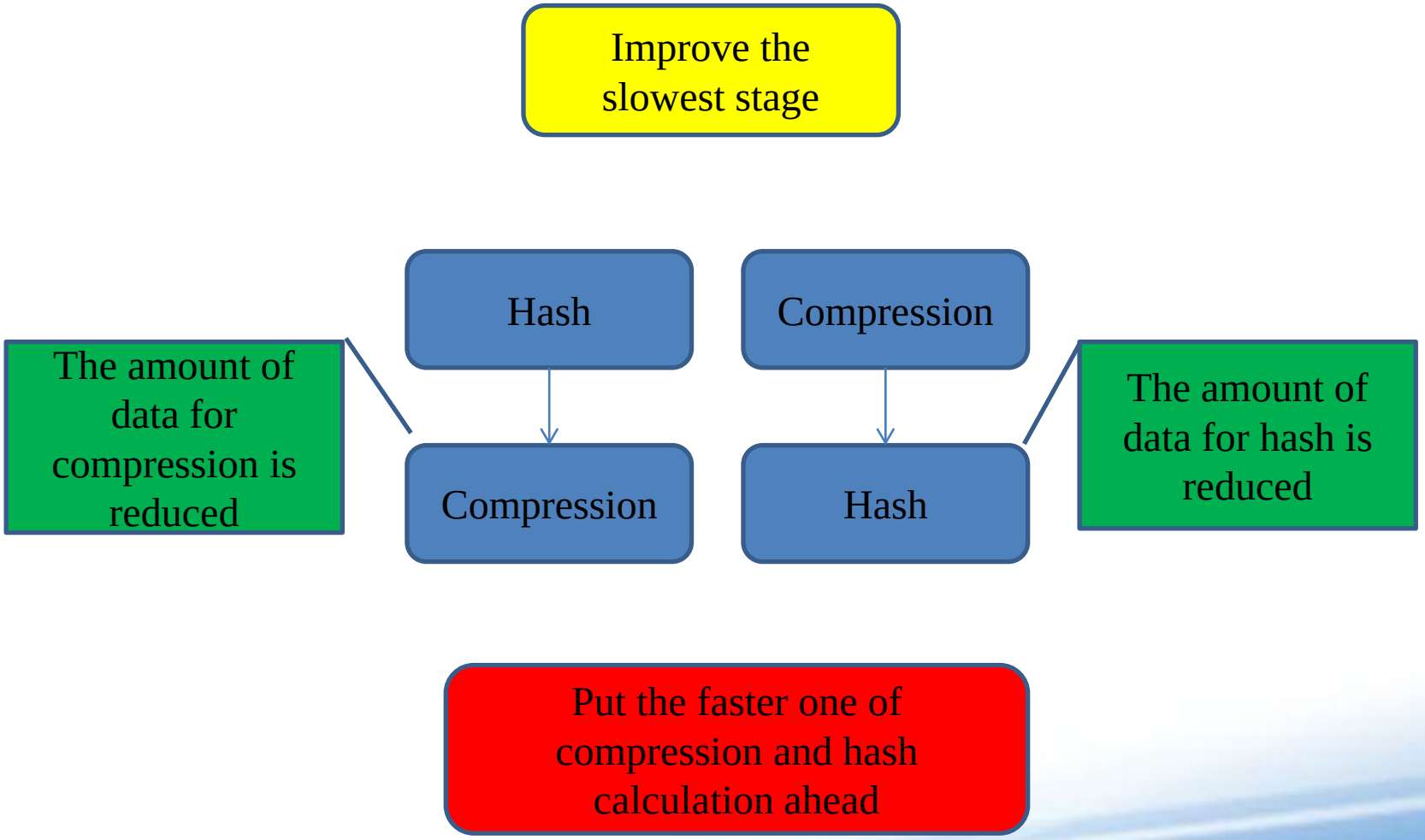
HICE

$$\frac{S \times (1 - D_r)}{C} + \frac{S}{H} + T_{Duplication\ identification} + T_{Encryption}$$

$$\frac{R_D H}{(1 - R_C)C} < 1$$

CHIE

Unbalanced Situation



Environment

- VIA
  - VIA Nano processor L2200@1600MHz
  - 2GB RAM, 2×32GB SSD
  - NVIDIA GTX 480, CUDA 3.0
- AMD
  - AMD Phenom(tm) II X4 745 Processor
  - 4GB RAM, 500GB 7200 rpm SATA disk
  - NVIDIA GTX 480, CUDA 4.0

Data

- MIRROR
  - Snapshot of Linux operating system
- SVN
  - Data from subversion server
- KERNEL
  - Linux kernel source code (Uncompressed)

Result

Platform	Data Type	Compression Throughput (MB/s)	Hash Throughput (MB/s)	R_C	R_D	$\frac{R_D H}{(1-R_C)C}$	CHIE Throughput (MB/s)	HICE Throughput (MB/s)
VIA	MIRROR	84.22	62.88	0.40	0.55	0.68	53.26	45.62
VIA	SVN	62.53	62.88	0.94	0.41	6.87	32.69	36.18
VIA	KERNEL	64.70	62.88	0.54	0.10	0.21	40.67	33.53
PadLock	MIRROR	84.22	299.72	0.40	0.55	3.26	71.49	96.28
PadLock	SVN	62.53	299.72	0.94	0.41	32.75	50.95	61.21
PadLock	KERNEL	64.70	299.72	0.54	0.10	1.01	54.94	54.69
VIA-GPU	MIRROR	172.22	62.88	0.40	0.55	0.33	71.15	50.77
VIA-GPU	SVN	121.29	62.88	0.94	0.41	3.54	39.95	42.87
VIA-GPU	KERNEL	161.74	62.88	0.54	0.10	0.08	58.85	44.20
PadLock-GPU	MIRROR	172.22	299.72	0.40	0.55	1.60	109.67	123.01
PadLock-GPU	SVN	121.29	299.72	0.94	0.41	16.89	73.48	83.99
PadLock-GPU	KERNEL	161.74	299.72	0.54	0.10	0.40	97.78	89.86
AMD	MIRROR	237.75	201.88	-	-	-	232.29	192.75
AMD	SVN	154.70	201.88	-	-	-	144.49	148.31
AMD	KERNEL	180.81	201.88	-	-	-	178.09	189.79
AMD-GPU	MIRROR	297.09	201.88	-	-	-	287.57	190.59
AMD-GPU	SVN	166.59	201.88	-	-	-	147.24	152.09
AMD-GPU	KERNEL	264.61	201.88	-	-	-	219.35	173.67

Platform	Data Type	Compression Throughput	Hash Throughput	
PadLock-GPU	SVN	121.29 MB/s	299.72 MB/s	
R_C	R_D	$\frac{R_D H}{(1-R_C)C}$	CHIE throughput	HICE throughput
0.94	0.41	16.89	73.48 MB/s	83.99 MB/s

Result

Platform	Data Type	Compression Throughput (MB/s)	Hash Throughput (MB/s)	R_C	R_D	$\frac{R_D H}{(1-R_C)C}$	CHIE Throughput (MB/s)	HICE Throughput (MB/s)
VIA	MIRROR	84.22	62.88	0.40	0.55	0.68	53.26	45.62
VIA	SVN	62.53	62.88	0.94	0.41	6.87	32.69	36.18
VIA	KERNEL	64.70	62.88	0.54	0.10	0.21	40.67	33.53
PadLock	MIRROR	84.22	299.72	0.40	0.55	3.26	71.49	96.28
PadLock	SVN	62.53	299.72	0.94	0.41	32.75	50.95	61.21
PadLock	KERNEL	64.70	299.72	0.54	0.10	1.01	54.94	54.69
VIA-GPU	MIRROR	172.22	62.88	0.40	0.55	0.33	71.15	50.77
VIA-GPU	SVN	121.29	62.88	0.94	0.41	3.54	39.95	42.87
VIA-GPU	KERNEL	161.74	62.88	0.54	0.10	0.08	58.85	44.20
PadLock-GPU	MIRROR	172.22	299.72	0.40	0.55	1.60	109.67	123.01
PadLock-GPU	SVN	121.29	299.72	0.94	0.41	16.89	73.48	83.99
PadLock-GPU	KERNEL	161.74	299.72	0.54	0.10	0.40	97.78	89.86
AMD	MIRROR	237.75	201.88	-	-	-	232.29	192.75
AMD	SVN	154.70	201.88	-	-	-	144.49	148.31
AMD	KERNEL	180.81	201.88	-	-	-	178.09	189.79
AMD-GPU	MIRROR	297.09	201.88	-	-	-	287.57	190.59
AMD-GPU	SVN	166.59	201.88	-	-	-	147.24	152.09
AMD-GPU	KERNEL	264.61	201.88	-	-	-	219.35	173.67

Platform	Data Type	Compression Throughput
AMD-GPU	MIRROR	297.09 MB/s
Hash Throughput	**CHIE throughput**	**HICE throughput**
201.88 MB/s	287.57 MB/s	190.59 MB/s

That’s all
Thank you!

Q&A