tar2h5: Small Files Packer for Machine Learning Tasks

Gerd Heber (The HDF Group), Dawei Mu (NCSA), Volodymyr Kindratenko (NCSA)

ILLINOIS NCSA | National Center for Supercomputing Applications

- NSF-funded IBM cluster for Deep Learning applications
 - 16x compute nodes,
 - 640x physical CPU cores,
 - 64x Nvidia V100 GPUs
 - 224 TB of All-Flash Storage

The Origin of Machine Name

- 2001: a space odyssey
- Early concept of an artificial intelligence system
- Didn't end well and we decided to give "him" a second chance



IBM POWER9 CPUs

- 14nm finFET semiconductor
- Stronger Thread Performance SMT
- POWER ISA 3.0 Architecture
- Enhanced Cache Hierarchy
- NVIDIA NVLink 2.0
- I/O System PCIe Gen4
- 2x 20 Cores with SMT4
 - Map to OS as 160 CPUs per node

25G (x24)	DDR4 PHY	DDR4 PHY
236 (224)	(2 channels)	(2 channels)
	Memory	
	CPU CPU Controller	CPU CPU CPU CPU
L2\$	L2\$	
	L3\$	L3\$ L3\$ ≚
L3s	L3\$	L3\$ L3\$
2 L2\$		
CPU CPU	CPU CPU	
CORECORE		CPU
PCIe Ctrl	Interconne SMP Ctrl	ct Accelerators 꽃
L3\$	L3\$	Accelerators L3\$ L3\$
L2\$	L25	L2\$ L2\$ WS
		The set was all and a set was all the set was all and and a set was all and and all all all all all all all all all al
		CPU
A S CORECORE	CORECORE Memory Controller	CORECORECORE
	DDR4 PHY	DDR4 PHY
25G (x24)	(2 channels)	(2 channels)

NVIDIA V100 GPUs

- Peak 7.8 TFLOP/s (double-precision).
- Peak 15 TFLOP/s (single-precision).
- SM / Cores : 80 / 5120.
- HBM2 Memory 16 GB : 900 GB/s.
- Config up to 128 KB L1 Cache per SM.
- Config up to 96 KB Shared Memory per SM.
- Constant memory 64 KB.
- 65536 32-bit Registers per SM.
- Clock Frequency : 1530 MHz

							L1 Instru	ctio	n Cache							
		L0 I	nstruc	tion C	ache						L0 li	nstruc	tion C	ache		
Warp Scheduler (32 thread/clk) Dispatch Unit (32 thread/clk)						Warp Scheduler (32 thread/clk)										
	Dis	spatc	h Unit	(32 th	read/c	lk)				Di	spatc	h Unit	(32 th	read/	clk)	
	Reg	ister	File (16,38	4 x 32	-bit)				Reg	jister	File ('	16,38	4 x 32	2-bit)	
FP64	INT	INT	FP32	FP32					FP64	INT	INT	FP32	FP32	H		
FP64	INT	INT	FP32	FP32					FP64	INT	INT	FP32	FP32	+		
FP64	INT	INT	FP32	FP32					FP64	INT	INT	FP32	FP32	H		
FP64	INT	INT	FP32	FP32	TEN	SOR	TENSOR		FP64	INT	INT	FP32	FP32	TEN	ISOR	TENSO
FP64	INT	INT	FP32	FP32	co	RE	CORE		FP64	INT	INT	FP32	FP32	cc	ORE	CORE
FP64	INT	INT	FP32	FP32					FP64	INT	INT	FP32	FP32			
FP64	INT	INT	FP32	FP32					FP64	INT	INT	FP32	FP32			
FP64	INT	INT	FP32	FP32					FP64	INT	INT	FP32	FP32			
LD/ LD/ ST ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	SFU		LD/ LD/ ST ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	SFU
	Mor		nstruc			olk)		ĬĒ		Wa		nstruc			(alk)	
		p Scl	nedule	r (32 t	hread/						rp Scł	nedule	r (32 t	hread		
	Dis	p Sch spatc	nedule h Unit	r (32 t (32 th	hread/ read/c	lk)				Di	rp Sch spatci	nedule h Unit	r (32 t (32 th	hread read/o	cik)	
	Dis Reg	p Sch spatc ister	nedule h Unit File ('	r (32 t (32 th 16,384	hread/	lk)				Di Reg	rp Sch spatc jister	nedule h Unit File ('	r (32 t (32 th 16,38	hread read/o	cik)	
FP64	Dis Reg INT	p Sch spatc ister INT	hedule h Unit File (* FP32	r (32 t (32 th 16,38 FP32	hread/ read/c	lk)			FP64	Di Reg INT	rp Sch spatc jister INT	nedule h Unit File (* FP32	r (32 t (32 th 16,38 FP32	hread read/o	cik)	
FP64	Dis Reg INT INT	p Sch spatc ister INT INT	File (FP32 FP32	r (32 th (32 th 16,38 FP32 FP32	hread/ read/c	lk)			FP64	Di Reg INT INT	rp Sch spatc jister INT INT	File (* FP32	r (32 t (32 th 16,38 FP32 FP32	hread read/o	cik)	
FP64 FP64	Dis Reg INT INT INT	rp Sch spatc ister INT INT	FP32 FP32 FP32 FP32	r (32 th (32 th 16,38 FP32 FP32 FP32	hread/ read/c	lk)			FP64 FP64	Di Reg INT INT INT	rp Sch spatc jister INT INT	File (* FP32 FP32 FP32 FP32	r (32 th (32 th 16,38 FP32 FP32 FP32	hread read/o	cik)	
FP64 FP64 FP64	Reg INT INT INT INT	rp Sch spatc ister INT INT INT	File (* File (* FP32 FP32 FP32 FP32 FP32	r (32 th (32 th 16,38 FP32 FP32 FP32 FP32	hread/ read/c	lk) -bit) SOR	TENSOR		FP64 FP64 FP64	Di Reg INT INT INT INT	rp Sch spatc jister INT INT INT	File (* File (* FP32 FP32 FP32 FP32 FP32	r (32 t (32 th 16,38 FP32 FP32 FP32 FP32	hread read/ 4 x 32 TEN	cik)	TENSOI
FP64 FP64 FP64 FP64	Dis Reg INT INT INT INT	p Sch spatc ister INT INT INT INT	File (* File (* FP32 FP32 FP32 FP32 FP32	r (32 th (32 th 16,38 FP32 FP32 FP32 FP32 FP32	hread/c read/c 4 x 32 TEN	lk) -bit) SOR			FP64 FP64 FP64 FP64	Di Reg INT INT INT INT	rp Sch spatc jister INT INT INT INT	File (* File (* FP32 FP32 FP32 FP32 FP32	r (32 t (32 th 16,38 FP32 FP32 FP32 FP32 FP32	hread read/ 4 x 32 TEN	2-bit)	
FP64 FP64 FP64 FP64 FP64	Dis Reg INT INT INT INT INT	p Sch spatc ister INT INT INT INT INT	File (* FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	r (32 th (32 th 16,38 FP32 FP32 FP32 FP32 FP32 FP32	hread/c read/c 4 x 32 TEN	lk) -bit) SOR			FP64 FP64 FP64 FP64 FP64	Di Reg INT INT INT INT INT	rp Sch spatc jister INT INT INT INT INT	FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	r (32 th (32 th 16,38 FP32 FP32 FP32 FP32 FP32 FP32	hread read/ 4 x 32 TEN	2-bit)	
FP64 FP64 FP64 FP64 FP64 FP64	Dis Reg INT INT INT INT INT INT	p Sch spatc ister INT INT INT INT INT	Fraz FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	r (32 th (32 th 16,38- FP32 FP32 FP32 FP32 FP32 FP32 FP32	hread/c read/c 4 x 32 TEN	lk) -bit) SOR			FP64 FP64 FP64 FP64 FP64 FP64	Di Reg INT INT INT INT INT INT	rp Sch spatc jister INT INT INT INT INT INT	File (* FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	r (32 t (32 th 16,38- FP32 FP32 FP32 FP32 FP32 FP32 FP32	hread read/ 4 x 32 TEN	2-bit)	
FP64 FP64 FP64 FP64 FP64	Dis Reg INT INT INT INT INT	p Sch spatc ister INT INT INT INT INT	File (* FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	r (32 th (32 th 16,38- FP32 FP32 FP32 FP32 FP32 FP32 FP32	hread/c read/c 4 x 32 TEN	lk) -bit) SOR	CORE		FP64 FP64 FP64 FP64 FP64	Di Reg INT INT INT INT INT	rp Sch spatc jister INT INT INT INT INT	FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	r (32 t (32 th 16,38- FP32 FP32 FP32 FP32 FP32 FP32 FP32	hread read/ 4 x 32 TEN	2-bit)	CORE
FP64 FP64 FP64 FP64 FP64 FP64 FP64	Dis Reg INT INT INT INT INT INT	P Scl spatc ister INT INT INT INT INT INT	Hedule h Unit File (FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	r (32 th (32 th 16,38 FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	hread/c 4 x 32 TEN: CO	sor RE			FP64 FP64 FP64 FP64 FP64 FP64 FP64	Di Reç INT INT INT INT INT INT	rp Sch spatc spatc int int int int int int int	File (' File (' FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	r (32 t (32 th 16,38- FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	hread read/u 4 x 32	2-bit)	
FP64 FP64 FP64 FP64 FP64 FP64 FP64 LD/LD/LD/	Dis Reg INT INT INT INT INT INT INT	P Scl spatc ister INT INT INT INT INT INT INT INT	File (File (FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	r (32 t (32 th 16,38 FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	hread/c read/c 4 x 32: TEN: CO	LD/	CORE		FP64 FP64 FP64 FP64 FP64 FP64 LD/ LD/ ST ST	Di Reg INT INT INT INT INT INT INT LD/ ST	rp Sch spatc ister INT INT INT INT INT INT INT INT INT	File (' FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	r (32 t (32 th 16,38 FP32 FP32 FP32 FP32 FP32 FP32 FP32 FP32	LD/	2-bit) 2-bit) SOR DRE	

DDN GS400NVE Flash Arrays Server

- 224 TB usable GPFS
- 8x EDR Inifiniband 100 GB/s bandwidth
- The mean time to list all home directories has been 7ms with a standard deviation of just 17.8%



HAL Software Overview

HAL Software

- OS : CentOS Linux 7.7
- Compilers :
 - GNU 4.8.5
 - Advance Toolchain 12.0
 - IBM XL 16.1.1
 - CUDA 10.2.89
 - PGI 2019.10
- Tools :
 - PowerAI 1.7.0 (Watson Machine Learning Community Edition)

- OpenMPI 4.0.3
- CMake 3.14.0
- Singularity 3.5.3

The IO Challenge

The catalyst of the rise of machine learning - Datasets

- dataset composed of millions of small files
- dominant random-access pattern

Researchers new to this area

- have expertise in domain science
- don't have a lot of HPC experience
- use small files as dataset
- produce overwhelming workload to shared storage
- => we used to reduce compute nodes to ease the I/O pressure

Tar2h5: Small Files Packer

Convert Tape ARchives to HDF5 files

- easy to use
- solutions for different scenarios

Functions

- archive checker
- h5compactor
- h5shredder



archive checker

- archive_checker
 - check how many files can be extracted from the input tar file.

- archive_checker_64k
 - check if any files within input tar files larger than 64 KB.

h5compactor

h5compactor

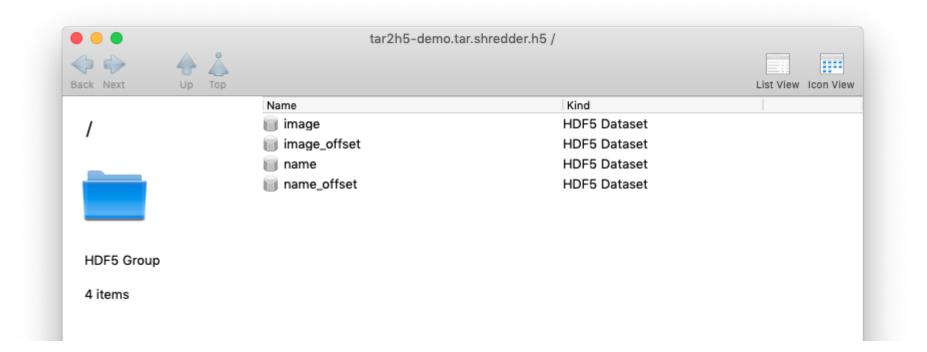
• converts input tar file to HDF5 file, all files within tar file should smaller than 64KB, using the file names as dataset names.

• • •	tar2h5-demo.tar.comp	actor.h5 /	
Gack Next Up T	20		List View Icon View
	Name	Kind	
1	📷 beverage-cafe.jpg	HDF5 Dataset	
/	📷 bokeh-cafe.jpg	HDF5 Dataset	
	📷 cappuccino-close-up.jpg	HDF5 Dataset	
	📊 coffee-beans-roasted.jpg	HDF5 Dataset	
	📊 coffee-shop.jpg	HDF5 Dataset	
	coffee-signage.jpg	HDF5 Dataset	
	coffeehouse.jpg	HDF5 Dataset	
HDF5 Group	up-of-coffee.jpg	HDF5 Dataset	
	espresso-machine.jpg	HDF5 Dataset	
10 items	wooden-table.jpg	HDF5 Dataset	

- h5compactor-sha1
 - converts input tar file to HDF5 file, all files within tar file should smaller than 64KB, using files' SHA-1 digest as dataset names.

	tar2h5-demo.tar.compactor-sha1.h5 /	
sack Next Up Top		List View Icon View
	Name Kind	1
1	📷 4aee075b287ed6eb5511a8e3cc684aab HDF5 Dataset	
,	🗑 5fee3ce79973fcdfc527676fdc0500e03e HDF5 Dataset	
_	同 6a4c01cd9553ee6a9138bbef07d6c413 HDF5 Dataset	
	🗑 9bb220c0fbf763e9317297b2ec8c54dff HDF5 Dataset	
	🗑 13f656934e35815b37a819989de902c7 HDF5 Dataset	
	🗑 567fc353709b3b076c292c0ade481224 HDF5 Dataset	
	同 011403426a1ad26ef72d6002de52a7f9f HDF5 Dataset	
HDF5 Group	🗑 74826935fd0d15258783e489e27fe805 HDF5 Dataset	
10 items	💼 abff3e94c328f8edae2b1b340b050a093 HDF5 Dataset	
	c23e360299ca363f10c3ca73b7e515ef9 HDF5 Dataset	

- h5shredder
 - converts input tar file to HDF5 file, *no file size limitation*, concatenate data, names, and offsets into 4 separate arrays for random access.





User Case

Project Title

• Efficient Large-Scale Video Generation with GANs

HAL User

• CS Research Assistant: Daniel B. McKee

Data Info

- large Kinetics-400 video dataset composed of around 240k videos
- the number of files is around 25 million
- total size of the JPG dataset around 125GB as a tar file

Performance Comparison

• Loading from the compact HDF5 file made training about 5x faster

tar2h5 Open Source

- The tar2h5 tools are available on GitHub
 - https://github.com/HDFGroup/tar2h5

→ C û 🔽 🔒 htt	ps://github.com/HDFGroup/tar2h 🚥 🗵 🛓		13 😁 🖸 🐀 🄅 🌻 1
Search or jump to	Pulls Issues Codespaces N	1arketplace Explore	¢ +• (
HDFGroup / tar2h5		⊙ Unwatch → 6	Star 0 Star 0
Code 🕙 Issues	김 Pull requests ⓒ Actions 때 Projects	🕮 Wiki 🕕 Secu	rity 🗠 Insights
🕈 master 👻	Go to file Add file -	🛓 👱 Code 🗸	About
gheber Added GPU dire	ct comments	on Sep 2 🕚 29	Convert Tape ARchives to HDF5 files
demo	remove temp files.	3 months ago	🛱 Readme
img	add sample images.	3 months ago	
src	Added IEEE FP16.	last month	Releases
gitignore	add ignore file	4 months ago	No releases published Create a new release
CMakeLists.txt	add test cases and README info.	3 months ago	
README.md	Update README.md	3 months ago	Packages
float-my-boat.org	Added GPU direct comments.	last month	No packages published
🖞 tar-to-hdf5.html	Minor touch-up.	4 months ago	Publish your first package
tar-to-hdf5.org	Forgot the Org file.	3 months ago	Contributors 2
🖞 tar2h5.html	Updated org file. Added (tangled) C source and (4 months ago	-
tar2h5.org	Fixed parameters.	4 months ago	gheber Gerd Heber
README.md		Ø	superdavidxp Dawei Mu
tar2h5			Languages
Convert Tape ARchives	to HDF5 files		• C 87.1% • CMake 12.9%
• archive_checker cl tar file.	neck how many files can be extracted from	the input	
 archive_checker_64 than 64 KB. 	k check if any files within input tar file	es larger	

Visualization with HDFCompass

<u>https://support.hdfgroup.org/projects/compass/</u>

•	tar2h5-demo.tar.comp	pactor.h5 /	
ء 🔶 🔶	۵.		
ck Next Up 1	Name	Kind	List View Icon View
/	📦 beverage-cafe.jpg	HDF5 Dataset	
/	bokeh-cafe.jpg	HDF5 Dataset	
	appuccino-close-up.jpg	HDF5 Dataset	
	coffee-beans-roasted.jpg	HDF5 Dataset	
	📊 coffee-shop.jpg	HDF5 Dataset	
	📷 coffee-signage.jpg	HDF5 Dataset	
	🝿 coffeehouse.jpg	HDF5 Dataset	
HDF5 Group	up-of-coffee.jpg	HDF5 Dataset	
	📷 espresso-machine.jpg	HDF5 Dataset	
10 items	📷 wooden-table.jpg	HDF5 Dataset	

Future Work

• Mixed Precision Support in HDF5

- IEEE FP16
- Google BFloat16
- NVIDIA TensorFloat (TF32)
- AMD FP24

THANK YOU FOR YOUR TIME !

