## Scalable Silicon Compute

NIPS 2017 Workshop on Deep Learning at Supercomputer Scale

# GRAPHCORE

simon@graphcore.ai

### Our comprehension and mechanization of intelligence is nascent

- best models and algorithms change every year
- increasingly heterogeneous composition of models
- huge appetite for more compute throughput

### At least until we understand intelligence better, we need machines which...

- exploit massive parallelism
- are agnostic to model structure
- have a simple programming abstraction
- are efficient for both learning and inference



Graphs can represent arbitrary model structure and expose compute parallelism.

We need to expose <u>a lot</u> of parallelism... O(1000) work items/processor x O(1000) processors/chip x O(1000) chips/system

In a parallel computer, communication is part of the computation. It's attractive to compile "communication kernels", as well as compute kernels.





## **Types of Parallelism**

- kernel partitioning
- pipelining
- parallel paths
- model clones sharing parameters, different data
- model clones with different parameters/cost functions/hyper-params, shared data
- multiple models of different structure
- ...



### Kernel partitioning exhausted?



Volta V100 fmac utilization on DeepBench





### Challenges for the Age of Parallel Processors

- Power
- Memory
- Abstraction

As parallelism goes up, local resources per processor go down.



### **Post-Dennard Silicon Scaling**

- Transistors / chip ~30% / year
- Compute / Watt ~15% / year

Big	Xeon	exemp	lars:

65n	Woodcrest X5160 2c			
0511	Conroe X3085 2c			
15n	Lynnfield X3470 4c			
4011	Nehalem X7550 8c			
20-	Westmere E7-8870 10c			
3211	SandyBridge E5-2690 8c			
00-	lvyBridge E7-2890v2 15c			
2211	Haswell E7-8890v3 18c			
14n	Broadwell E7-8894v4 24c			
	Skylake 8180P 28c			



### Silicon efficiency is the full use of available power

- Keep data local
- Serialise communication and compute
- Re-compute what you can't remember

Proximity of memory is defined by energy, more than latency.





(rough metrics for 16nm)



All logic chips are power limited



Processor + memory systems @ 240W (for 300W card)

DRAM on interposer 180W GPU + 60W HBM2



16GB @ 900GB/s

Distributed SRAM on chip 120W IPU x2



600MB @ 90TB/s



## "Colossus" IPU

(in honour of Tommy Flowers)

- All-new pure distributed multi-processor for MI.
- Mostly memory, for "model on chip".
- A cluster of IPUs acts like a bigger IPU.
- Bulk Synchronous Parallel (BSP) execution.
- Stepwise-compiled deterministic communication.
- Programmed using standard frameworks (TensorFlow, ...) over Poplar<sup>™</sup> native graph abstraction.

	supertile tile frame	supertile.			supertile .	supertie	supert <i>le</i>	supertile tile frame	
da	: _superble : :	supertile;	supertile : :	supertile;	: :supertile : :	supertie :	Consultation Consu	: : : supertile :	E.
	supertile	superble	supertile	superties	supertile	supertie	supertie	supertie :	
aroe	subortio	striartie	supertile	A nertile	subditie	tile frame	sunditie	i - tie frame	
		tile.frame	tile frame		tile.frame	net tile frame	tile frame	tie frame	
	tie frame	superble	supertile tile frame	en utile frame	supertile	supertie	supertie	supertie	말끈
-47 E	si supectie : :	supertile.	supertile inte	this frame	supertile	supervie and	supertie	set i supert lé	
anae.	a superble	superble	supertile	supertile	supertile	supertie	supertie	supertie	10.00
-	supertile	supertile.	supertile	supertile.	pervtile.frame	supertie	supertie	tie frame	
	🚓 tie frame 💠	je	🚓 tile frame 💠	in tile frame:	🚌 bleiframe	tie die frame	e⇒.tl€frame	: ⇔tie frame	
aroe.	tie frame	supercie	tile_frame	- tile frame	tile.frame	- supertie	tie frame	tie frame	
E.	superble	ula finda	supertile	supertile	Supertile	supertie	supertie	supertie	LA
	supertile :	supertile;	j. jsupertile :	supertile;	supertile :	supertie :	supertile	upert le	
- 9-		supertile	s. supertile	set supertile	tile.frame	supertie	supertie	tie frame supertile	
ame	🚎 i tle framé 🗌		🚌 – tile Trame	initiation in the second				i 🚌 – i têc fraimé	ه نی ا
	tio frame	;supercis	dischering	) tile frame:	tile:frame	tile frame	tie frame	tie frame	
	n_supertile	superble tile frame	supertile	supertile tile frome	supertile tile frame	supertie	Supertie	supertile	Þ.,
aroe	supectile		supertile	supertile.	supertile .	supertie	súpértée	:supertilé :	
	supertile	supertile	supertile	superfile	supertile	supertie	supertie	supertile	- 1
	- tie frame	tile frame	tile frame	per stile frame		- stie frame	tie.frame	- parts, tile frame	
	supertile	supertile.	supertile	supertile.	supertile	supertie	supertie	supertile	
	supertile :	supertile	supertile :	Supertile	supertile	supertie	supertie :	supertile :-	
staroe.	superble	superble	supertile	superties	supertile .	supertie	supertie	· · · supertie	
	ni tie frame	et utile frame.	Hereit die Frame			. per tile frame.	- tieframe : .	i - tie framé	
	- the frame		- tile frame	- tile frame:		net tile frame	- tie frame	- tie frame	
		column_mux			.,.columri_mux	¢alµmn_múx			
	us ditrinte								
		- ux_1 frame	us_2 frame	, lac_3,fram¢	in uk_8 frame:	,uK_S;frame	r	Sus_7/rame	
	cupertile	ux_1.frame		supertiener	uk Bitarre	supertie	superferme		
	supert.le		Luk_2 frame	supertile-	supertile-and supertile-and supertile-and supertile-and	ux_S;framp	supertie	Lue_7 frame Lue_7 frame Lue_rame tile_frame Supertile_frame	
	supertile	ux_1.framd	supertile-ette 	supertile			supervie supervie supervie te frame supervie		<u>y</u> ]]
	supertile tile frame supertile tile frame supertile			superties	superfileframe superfilef ble.frame superfilef ble.frame superfilef	auc Siframe	supervie	uk_7/rama upertia tie frame upertia tie frame upertia	
	eupertile tie frame tie frame tie frame tie frame tie frame tie frame		Luc_2 frame		upartile- 		supertie supertie tie.frame supertie tie.frame supertie tie.frame tie.frame	us 7 frame supertile tie trans supertile supertile tie trans tie trans supertile tie trans	
		superties superties superties superties superties superties superties superties superties superties superties superties superties superties superties superties superties	Lug_2 frame			supertie- 		us 7 frame supert le- Ne trans Jupert le- supert le- supert le- tie trans the trans the trans the trans the trans the trans	- Internet and the second s
	. supertile- 	supertie- superi	Luc_2 frame	.upertie- .tile frame .tile frame	- supertile- -	supertie- - supertie- - tie frame - supertie- - tie frame - supertie- - tie frame - supertie- - supertie- - frame - supertie- - supertie- - supertie-	- us strane - useriare - useriare	supertile- tile trains Jupertile- tile trains Supertile- tile trains Supertile- Supertile- Supertile- Tile trains Supertile- Su	
	supertie	supervise-off su	us_2 frame	upertie upe	superile	Juperta Juperta Juperta Juperta Juperta Juperta Juperta Juperta Juperta Juperta Juperta Juperta Juperta Juperta Juperta Juperta Juperta	superfe- beframe superfe- superfe- beframe beframe beframe superfe- beframe superfe- beframe beframe beframe	uu_7/fromo	
	- cupartile	upertile- s	Le 2 france 	Les 3 frame 	ude 3 frame Bille frame Sille	uc Strang super to	Use frame superfermine -superfermi		
	oppentle=# the trans upertle=# the trans upertle=# the trans upertle=# the trans upertle=# the trans upertle=# the trans upertle=# upertle=# upertle=# upertle=#	-upartia- -upartia-	Le 2 frane uoentie-e-i uoentie		de 3 frans 	Jung Shamp Ling S	da Brans - Superla- - Superl	- ut_7 famp - ut_7 famp - uppertie- - the presence - the presence - uppertie- - uppertie-	
	usertie	- upartite	Lucational Lucational	Les 3 frame supertieres super	de 3 frans 		de Brans Superle- He Fame Superle- He Fame Superle- He fame Superle- He fame Superle- He fame Superle- He fame Superle- He fame Superle- He fame Superle- He fame Superle- Superle- He fame Superle- Superle- He fame Superle- Superle- He fame Superle- Superle- He fame Superle- Superle		
	uper le- tic frame yuper le- He frame super le- tic frame yuper le- yuper le- tic frame yuper le- ti	- up at Jrans - up a	Logitaria	- Juc 3 frame - Suprifier- Sile Fame - Suprifier- Sile Fame - Suprifier- - Supri	- stopertile- - stop	. Superior - Starts . Superio	44 Brans - Superior Herrane - Superior - Superior	supertiere supertiere The forms Use forms Use forms Supertiere Sup	
	-upertie	La franci La franci La franci Se fra	Lie Strans	- Jor 3 frame	- Lupertie- - Lupertie- Biel France - Lupertie- - Lup	Luc Sharno	Lie Strans Liefrans		
	upertile=" the hane _upertile=" the hane	- up 1 frand seperators - up and e	Us_2 frame	Juc 3 frame 		Superior to stars	- Superior Services - Superio	- ul. 7 farro - suparti - farro - suparti - farro - uparti - farro - uparti - farro - uparti - farro - uparti - farro - suparti - s	
	-upartie tie krane Juartie ke frane Juartie ke frane Juartie tie krane Juartie tie krane Juartie tie krane Juartie tie krane Juartie tie krane Juartie tie krane Juartie Juartie	- up 1 frand	Us a frame Lis promo- Lis promo- Lis promo- us frame - userfile- - userfile-	- Jac 3 frame	- Lugertie- - Lug	Langer time Langer time Lange	Line Johnson		
	Louantle	- up 1 frams - up art 0 error - up art 0 error	U.S. France	- Joe 2 Stante - Support Con- - Support Con-	- de a france - de a france - de a france - de a france - de grant - de g	See 5 Starts	- Les Diferies - Segler Roman La Segler Roman - Segler Roma	- Sul 7 forms - Superior - Superior - Su	
	Conservations Conser	- up at Drama - up a	U.S. Argense 	- Mayor Garane	Lipe 3 frants Billio frants - Liperitie -		Le La France He fanne Le	- Set 7 5 5000	
	- Superflow - Sup	- up 1 frame - up art 0	La provincia de la provincia d	La particular segnification de la construction de la constructi	superior of the second	Sectors and a sector and a sect	The set of terms of the set of terms of	a particular de la conservación	
	The banks of the second	superline of the second	La particular de la companya de la c	superformance in a strandormal in trans. Superformance in the strandormal internationa	Lange of the second sec	Superline of Super	Line of Stars	Jupite of Stars	
	Constraints of the second	Control of the second s	Lange Shore	and a second sec	Lange of Starst	see 1 See 1 See 1	Line of the second seco	searcher search	
	Annual and a second sec	A sent of the sent	Augusta in the second s	Construction C	search forms searcher of the searcher searcher of the searcher searcher searcher of the searcher searcher of the searcher searcher searcher of the searcher searcher o	superface de la forma de la construcción de la cons	La grande de ministration de la construcción de la	supplier in two supplier in two supplier in two supplier in two supplier in two supplier suppli	
	Constraints Constrain	- Long the	Lange Shores Lange Shores La	and a second sec	Lange de la forme la generation de la construction la construction de la construction de la la construct	sandar sandar	Line of the second seco	support sup	
			A spartie of the offer of the second	La contra de la co	scandar forms scandar and scandar and sca	superface in a superface in a superface in the superface	search of the se	Lange of the second sec	
	Comparison of the second	- Superitorial Sup	Lange of the second sec	and a second sec	Lange of the second sec	sandar sa	speed of streng inclusion of the strength inclusion of the strength in	supplier of the second	



#### Colossus pair on a PCIe card



all-to-all exchange spines each ~8TBps link + host bandwidth 384GBps/chip



### **Bulk Synchronous Parallel**



### **BSP** Execution Trace

![](_page_15_Figure_1.jpeg)

![](_page_15_Picture_2.jpeg)

### BSP Trace: ResNet-50 training, batch=4

![](_page_16_Figure_1.jpeg)

![](_page_16_Picture_2.jpeg)

## Managing Memory

IPU has smaller memory than GPU, but greater effective compute:

- Use small batches per IPU.
- Trade compute for memory.
- Use memory-efficient algorithms/models

An *efficient* gradient learner will converge in a minimum number of parameter updates ...so smaller batches should learn faster, as well as generalizing better. Efficient small batch machines allow learning to parallelise over more machines.

![](_page_17_Picture_6.jpeg)

### Re-compute what you can't remember

DenseNet-201 training, batch=16.

Naive strategy: memorize activations only at input of each residue block, recompute all others in backward pass.

Executing on CPU, recording total memory allocated for weights + activations. Float16 weights and activations, single weight copy.

![](_page_18_Figure_4.jpeg)

C NIPS 2017

### **Reversible Nets**

arxiv.org/abs/1707.04585

![](_page_19_Figure_2.jpeg)

No need to save any activations, except for layers which lose information (pool, stride).

![](_page_19_Picture_4.jpeg)

Some preliminary IPU benchmarks...

- ResNet-50 training at 2,250 images/s on 1 card with batch=8 16,000 image/s over 8 cards with batch=64
- DeepBench LSTM inference (per layer, 1536 hidden units, 50 steps) 60,000 iteration/s on 1 card at 7ms latency
- 600 full WaveNet voice generators on 1 card at 16k sample/s (MOS 3.35, 20 layers, 64 residual channels, 128 skip channels)

More at www.graphcore.ai

![](_page_20_Picture_5.jpeg)

![](_page_21_Picture_0.jpeg)

0c