Title: Tensor Processing Units (TPUs) as scientific supercomputers

Speakers: Guifre Vidal

Series: Quantum Matter

Date: October 06, 2022 - 3:00 PM

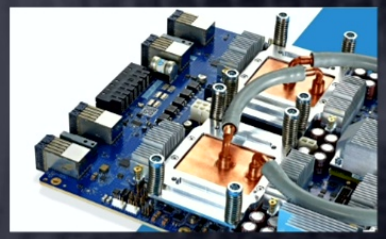
URL: https://pirsa.org/22100100

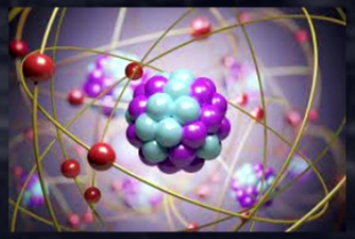
Abstract: Google's TPUs were exclusively designed to accelerate and scale up machine learning workloads, amid the ongoing planet-wide race to build faster specialized hardware for artificial intelligence. But one must surely be able to use this hardware for other challenging computational tasks, right? We explored how to turn a TPU pod (2048 TPU v3 cores) into a dense linear algebra supercomputer to e.g. multiply two matrices of size 1,000,000 x 1,000,000 in just 2 minutes. We then used this power to perform a number of quantum physics and quantum chemistry computations at scale. For instance, we recently completed two largest-ever computations: a Density Functional Theory DFT computation of electronic structure (with N = 248,000 orbitals), and a Density Matrix Renormalization Group DMRG computation (with bond dimension D = 65,000). Cloud-based TPU pods and GPU pods are accessible to anyone and are poised to revolutionize the scientific supercomputing landscape.

Zoom link: https://pitp.zoom.us/j/97006251134?pwd=WHhLa2pRUno0LzJjbzVnL2tsdGhOUT09

PERIMETER INSTITUTE ***Quantum Matter seminar*** Oct 6th 2022

Tensor Processing Units (TPUs) as scientific supercomputers





Guifre Vidal

Google Quantum AI

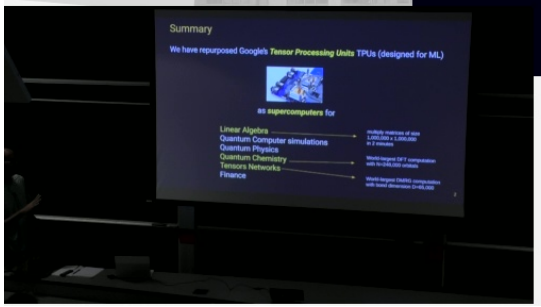
Summary

We have repurposed Google's ***Tensor Processing Units*** TPUs (designed for ML)

as ***supercomputers*** for

Linear Algebra →	multiply matrices of size 1,000,000 x 1,000,000 in 2 minutes
Quantum Computer simulations	
Quantum Physics	
Quantum Chemistry →	World-largest DFT computation with N=248,000 orbitals
Tensors Networks →	World-largest DMRG computation with bond dimension D=65,000
Finance	

2



Series of recent papers describing our work

A. Morningstar, M. Hauru, J. Beall, M. Ganahl, A. GM Lewis, V. Khemani, G. Vidal,
Simulation of quantum many-body dynamics with Tensor Processing Units: Floquet prethermalization, **arXiv:2111.08044** (PRX Quantum)

M. Hauru, A. Morningstar, J. Beall, M. Ganahl, A. Lewis, G. Vidal,
Simulation of quantum physics with Tensor Processing Units: brute-force computation of ground states and time evolution, **arXiv:2111.10466**

A. GM Lewis, J. Beall, M. Ganahl, M. Hauru, S. Basu Mallick, G. Vidal,
Large Scale Distributed Linear Algebra With Tensor Processing Units, **arXiv:2112.09017** (PNSA)

R. Pederson, J. Kozlowski, R. Song, J. Beall, M. Ganahl, M. Hauru, A. GM Lewis, S. Basu Mallick, V. Blum, G. Vidal,
Tensor Processing Units as Quantum Chemistry Supercomputers, **arXiv:2202.01255**

R. Shillito, A. Petrescu, J. Cohen, J. Beall, M. Hauru, M. Ganahl, A. G.M. Lewis, G. Vidal, A. Blais,
Dynamics of Transmon Ionization, **arXiv:2203.11235** (Physical Review Applied)

M. Ganahl, J. Beall, M. Hauru, A. G. M. Lewis, J. H. Yoo, Y. Zou, G. Vidal,
Density Matrix Renormalization Group with Tensor Processing Units, **arXiv:2204.05693**

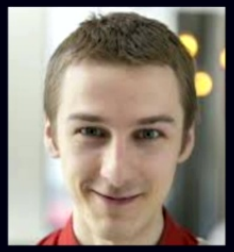
J. Kozlowski et al.,
Full Protein Density Functional Theory with Tensor Processing Units, (in preparation)

R. Song et al.,
Accelerating Quantum Chemistry Computations with Tensor Processing Units: From Biology to Materials Science, (in preparation)

Simulation & Optimization team within Sandbox@Alphabet
(rip Dec 2021)

Martin Ganahl

Research Scientist

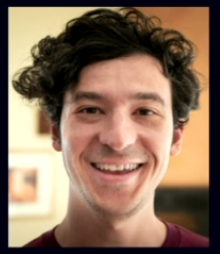


Markus Hauru

Research Scientist

Adam Lewis

Research Scientist



Jackson Beall

Software Engineer

Ryan Pederson

Resident Research Scientist (UCI)

Ruyi Song

Resident Research Scientist (Duke)

Johnny Kozlowski

Resident Research Scientist (UCI)

Guifre Vidal

Research Scientist Research Lead

Shrestha Basu Mallick

Product manager

Matthias Tan

Strategy & Partnerships

Olivier Lacombe

Strategy & Product

Stefan Leichenauer

Eng Lead

Jack Hidary

Director

4

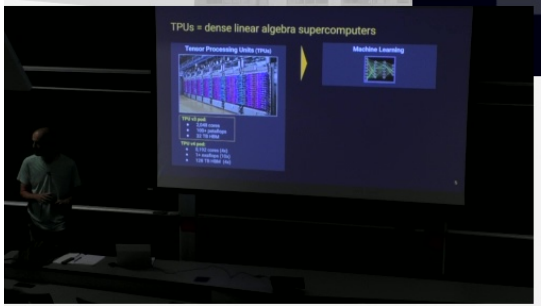
TPUs = dense linear algebra supercomputers

Tensor Processing Units (TPUs)

TPU v3 pod:
- 2,048 cores
- 100+ petaflops
- 32 TB HBM

TPU v4 pod:
- 8,192 cores (4x)
- 1+ exaflops (10x)
- 128 TB HBM (4x)

Machine Learning



TPUs = dense linear algebra supercomputers

Tensor Processing Units (TPUs)

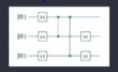
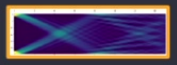
TPU v3 pod:
- 2,048 cores
- 100+ petaflops
- 32 TB HBM

TPU v4 pod:
- 8,192 cores (4x)
- 1+ exaflops (10x)
- 128 TB HBM (4x)

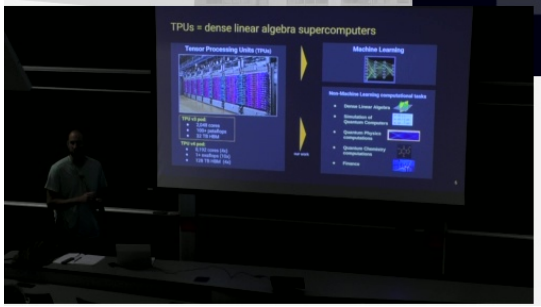
our work

Machine Learning

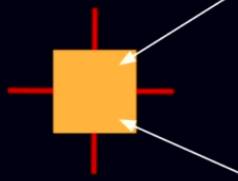
Non-Machine Learning computational tasks

- Dense Linear Algebra 
- Simulation of Quantum Computers 
- Quantum Physics computations 
- Quantum Chemistry computations 
- Finance 

6



Dense Linear Algebra Supercomputing. Why TPUs?



TPU v3 core

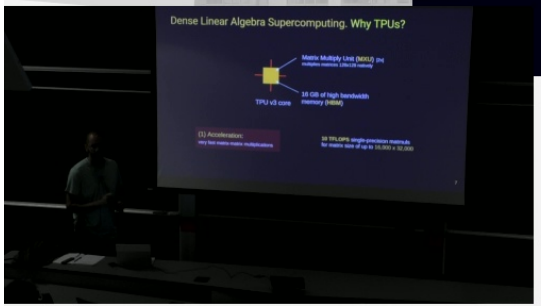
Matrix Multiply Unit (**MXU**) [2x]
multiplies matrices 128x128 natively

16 GB of high bandwidth memory (**HBM**)

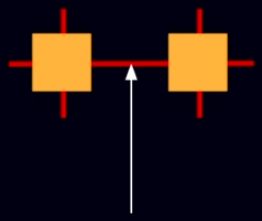
(1) Acceleration:
very fast matrix-matrix multiplications

10 TFLOPS single-precision matmuls
for matrix size of up to 16,000 x 32,000

7

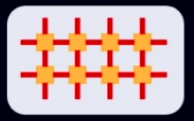


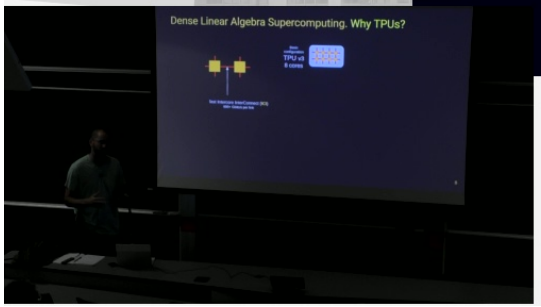
Dense Linear Algebra Supercomputing. Why TPUs?



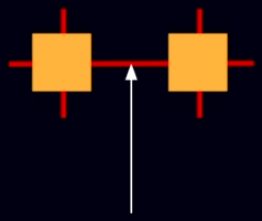
fast Intercore InterConnect (**ICI**)
600+ Gbits/s per link

Basic configuration
TPU v3
8 cores





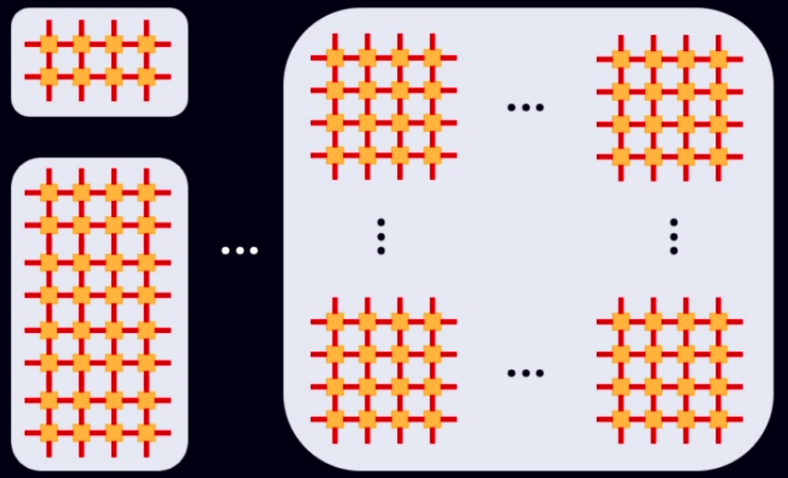
Dense Linear Algebra Supercomputing. Why TPUs?



fast Intercore InterConnect (**ICI**)
600+ Gbits/s per link

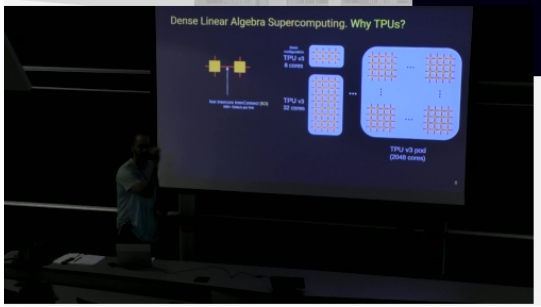
Basic configuration
TPU v3
8 cores

TPU v3
32 cores

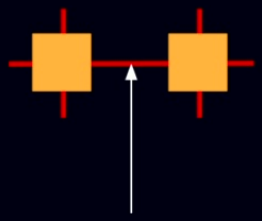


TPU v3 pod
(2048 cores)

8



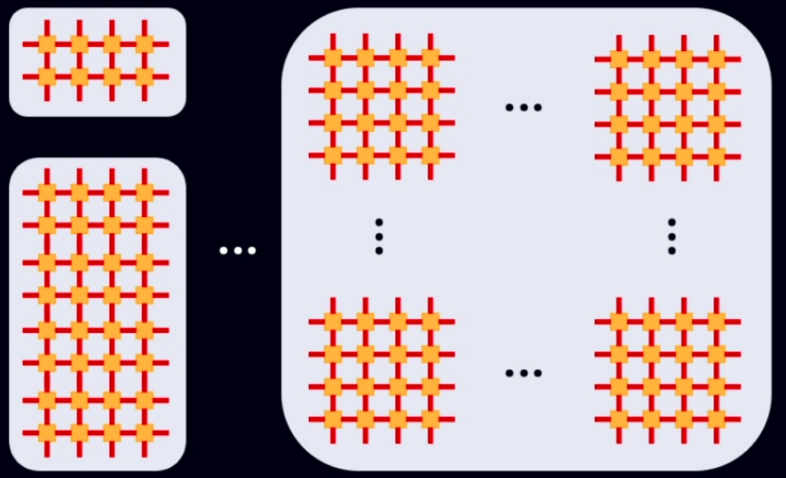
Dense Linear Algebra Supercomputing. Why TPUs?



fast Intercore InterConnect (**ICI**)
600+ Gbits/s per link

Basic configuration
TPU v3
8 cores

TPU v3
32 cores

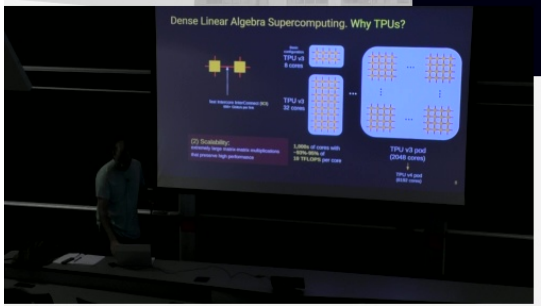


TPU v3 pod
(2048 cores)

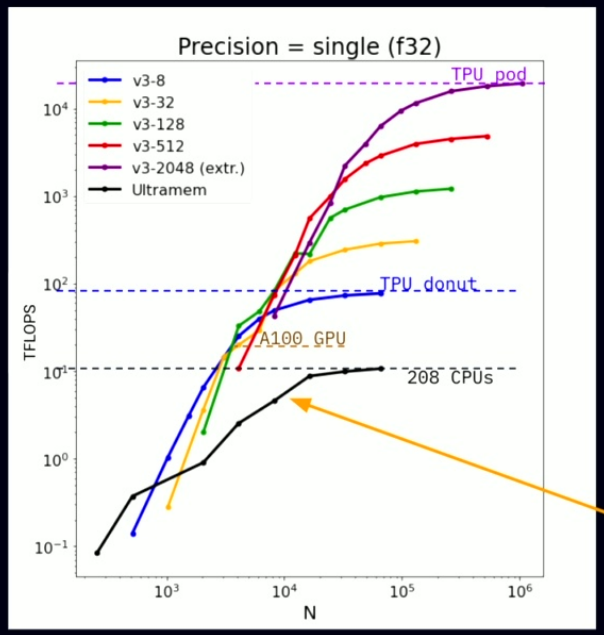
TPU v4 pod
(8192 cores)

(2) Scalability:
extremely large matrix-matrix multiplications that preserve high performance

1,000s of cores with
~93%-95% of
10 TFLOPS per core



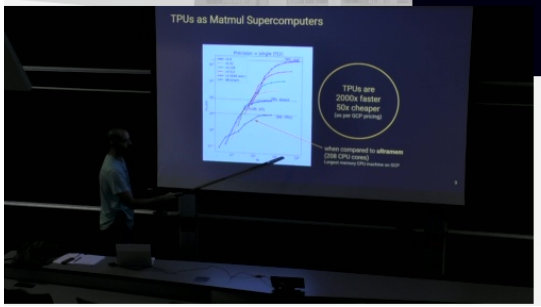
TPUs as Matmul Supercomputers



TPUs are
2000x faster
50x cheaper
(as per GCP pricing)

when compared to ***ultramem***
(208 CPU cores)
Largest memory CPU machine on GCP

9



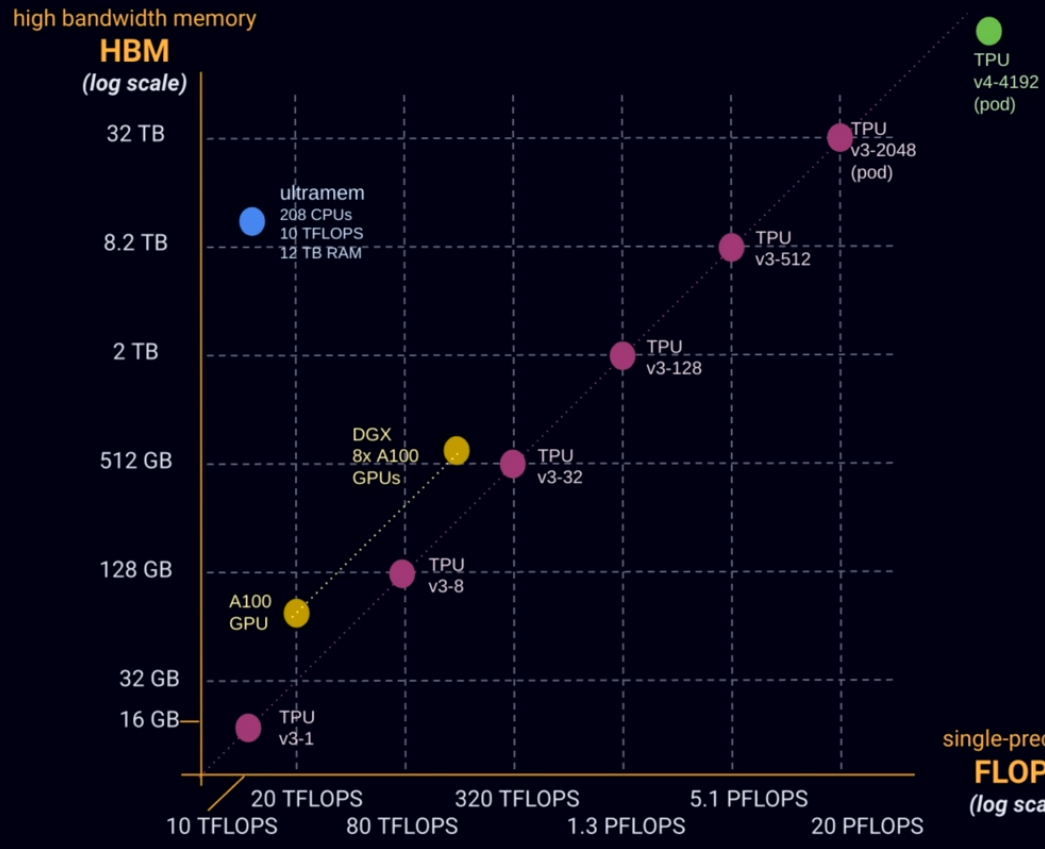
TPUs vs GPUs

for distributed matmuls
(single-precision)

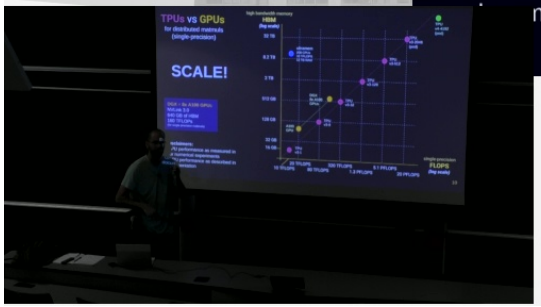
SCALE!

DGX = 8x A100 GPUs
NVLink 3.0
640 GB of HBM
160 TFLOPs
(for single-precision matmuls)

Disclaimers:
-TPU performance as measured in our numerical experiments
-GPU performance as described in documentation



10



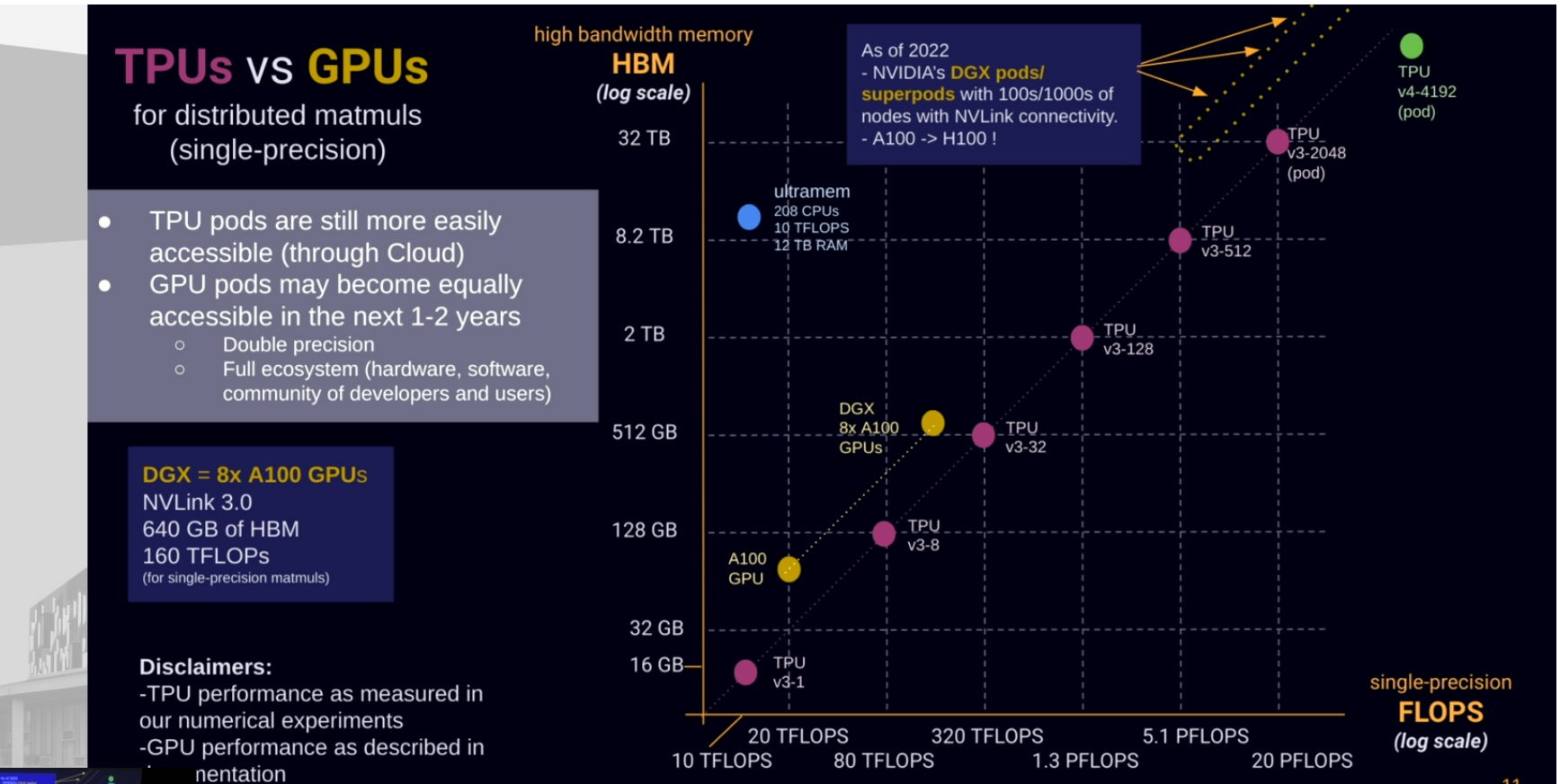
TPUs vs GPUs

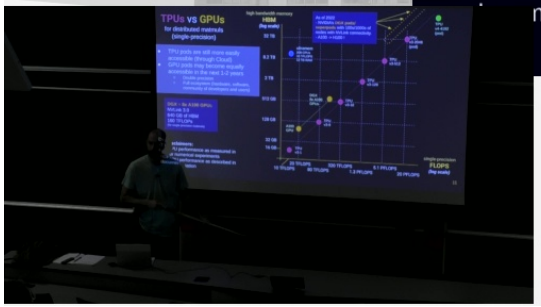
for distributed matmuls
(single-precision)

- TPU pods are still more easily accessible (through Cloud)
- GPU pods may become equally accessible in the next 1-2 years
  - Double precision
  - Full ecosystem (hardware, software, community of developers and users)

DGX = 8x A100 GPUs
NVLink 3.0
640 GB of HBM
160 TFLOPs
(for single-precision matmuls)

Disclaimers:
-TPU performance as measured in our numerical experiments
-GPU performance as described in
nentation





Two flavours of (distributed, dense) matrix multiplication

1) *square x square = square*

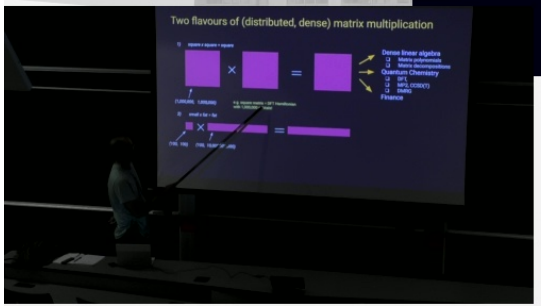
(1,000,000, 1,000,000)

e.g. square matrix = DFT Hamiltonian with 1,000,000 orbitals!

- Dense linear algebra
  - ❑ Matrix polynomials
  - ❑ Matrix decompositions
- Quantum Chemistry
  - ❑ DFT,
  - ❑ MP2, CCSD(T)
  - ❑ DMRG
- Finance

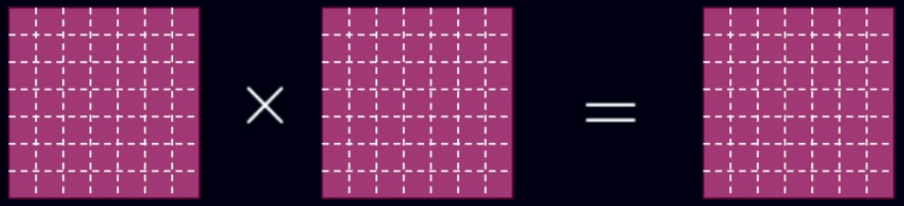
2) *small x fat = fat*

(100, 100) *(100, 10,000,000,000)*



Two flavours of (**distributed**, dense) matrix multiplication

1) *square x square = square*

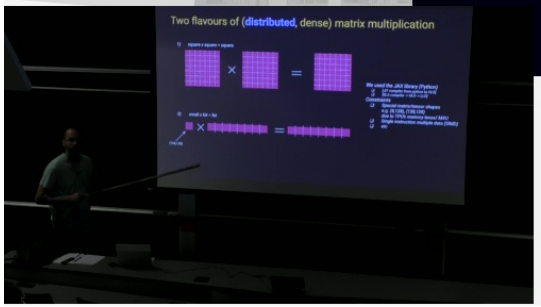


2) *small x fat = fat*

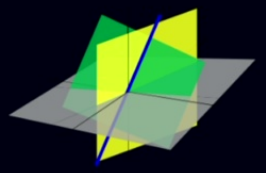
(128,128)

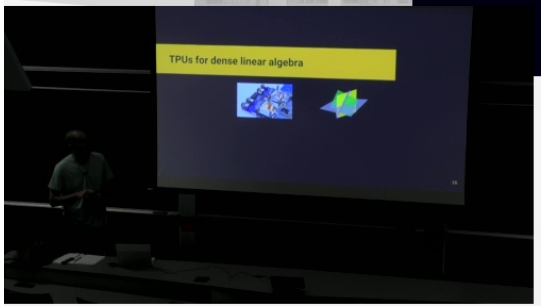
We used the JAX library (Python)
- *[JIT compiler from python to HLO]*
- *[XLA compiler -> HLO -> LLO]*

Constraints
- *Special matrix/tensor shapes*
  e.g. (8,128), (128,128)
  due to TPU's memory lanes/ MXU
- *Single instruction multiple data (SIMD)*
- *etc*



TPUs for dense linear algebra





Distributed dense linear algebra

TPUs can accelerate and scale up matrix multiplications (for dense matrices)

→ 0ther linear algebra operations?

Examples:

- Polar decomposition
- Sign function
- Inverse square root
- Singular value decomposition

needed for quantum chemistry

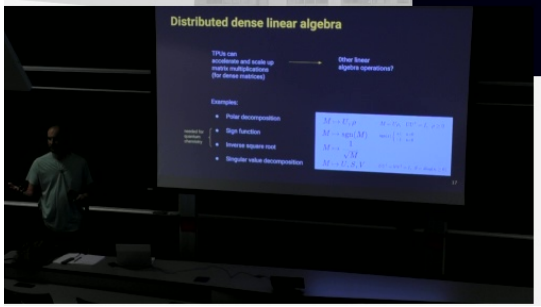
$$M \mapsto U, \rho \qquad M = U\rho, \quad UU^\dagger = I, \quad \rho \geq 0$$

$$M \mapsto \mathrm{sgn}(M) \qquad \mathrm{sgn}(x) \begin{cases} +1 & \mathrm{x}>0 \\ -1 & \mathrm{x}<0 \end{cases}$$

$$M \mapsto \frac{1}{\sqrt{M}}$$

$$M \mapsto U, S, V \qquad UU^\dagger = VV^\dagger = I, \quad S = diag(s_i \geq 0)$$

17



Polar decomposition

$$M \mapsto U, \rho \qquad M = U\rho, \quad UU^\dagger = I, \quad \rho \geq 0$$

How?

$$X_{k+1} = \frac{1}{2} X_k (3I - X_k^\dagger X_k)$$

$$X_0 = M$$

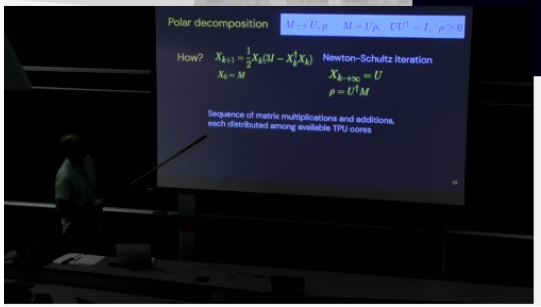
Newton–Schultz iteration

$$X_{k\to\infty} = U$$

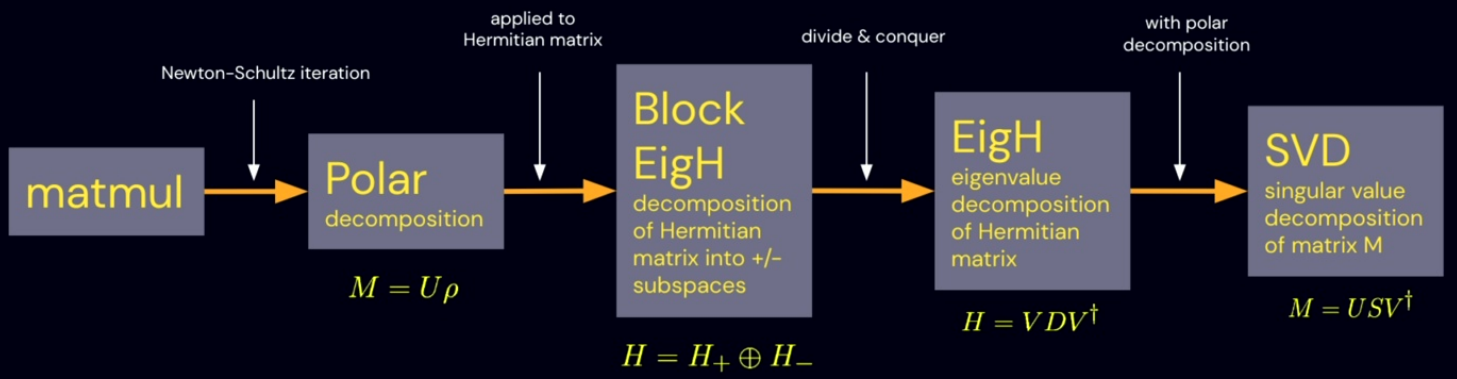
$$\rho = U^\dagger M$$

Sequence of matrix multiplications and additions,
each distributed among available TPU cores

18



Singular Value Decomposition



technical guide

projectors onto positive/ negative eigen-subspaces

$$P_\pm = \frac{I \pm U}{2}$$

$$H_+ = P_+ H P_+ > 0$$

$$H_- = P_- H P_- < 0$$

Apply Block EigH recursively

Example: define

$$H' = H_+ - \sigma I$$

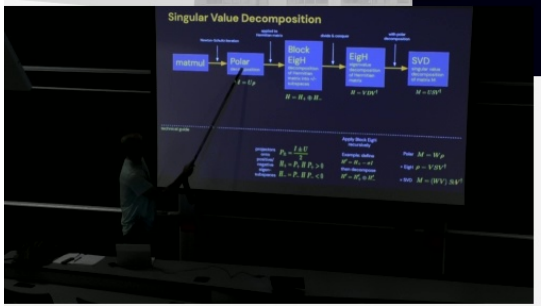
then decompose

$$H' = H'_+ \oplus H'_-$$

Polar $M = W\rho$

+ EigH $\rho = VSV^\dagger$

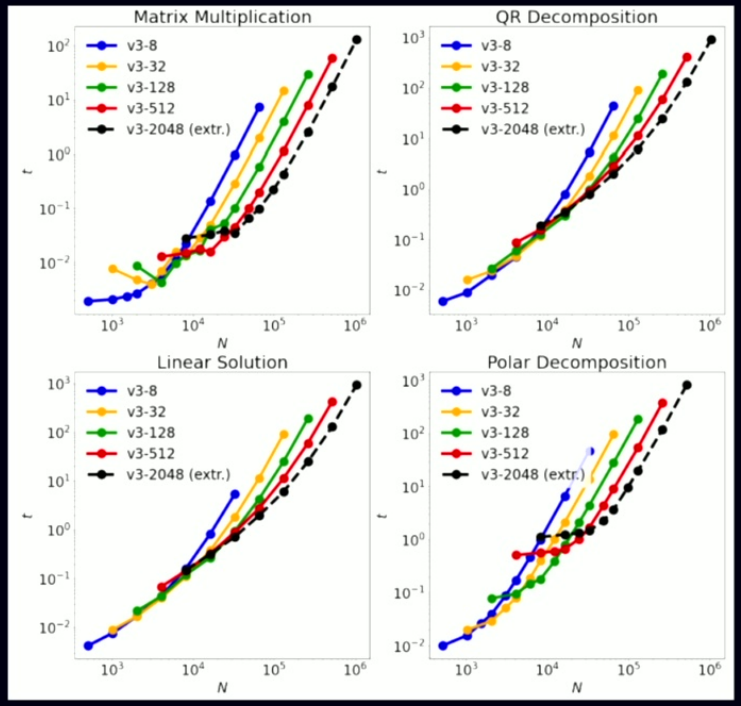
= SVD $M = (WV)\,SV^\dagger$

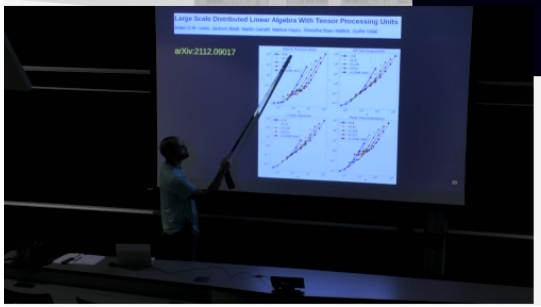


Large Scale Distributed Linear Algebra With Tensor Processing Units

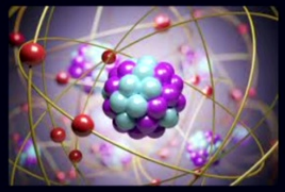
Adam G.M. Lewis, Jackson Beall, Martin Ganahl, Markus Hauru, Shrestha Basu Mallick, Guifre Vidal

arXiv:2112.09017

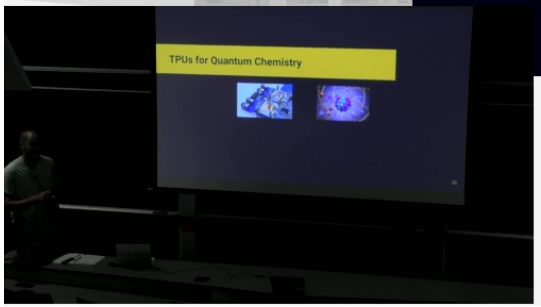




TPUs for Quantum Chemistry



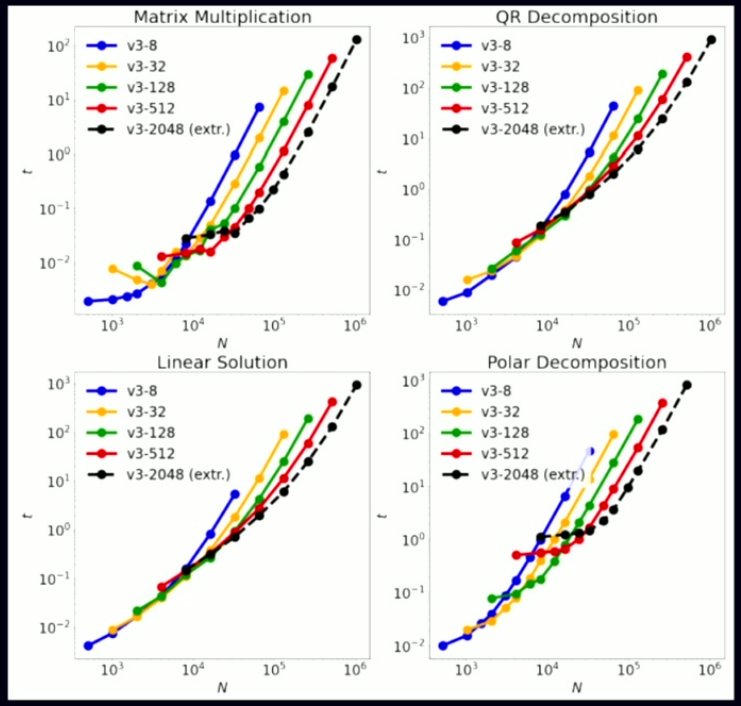
21

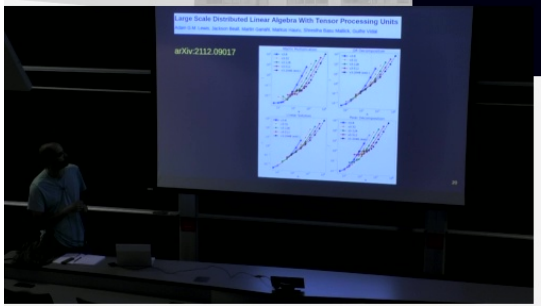


Large Scale Distributed Linear Algebra With Tensor Processing Units

Adam G.M. Lewis, Jackson Beall, Martin Ganahl, Markus Hauru, Shrestha Basu Mallick, Guifre Vidal

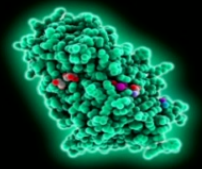
arXiv:2112.09017





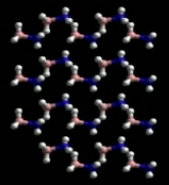
Quantum Chemistry Applications

Nucleic acids
(DNA, RNA, TNA)

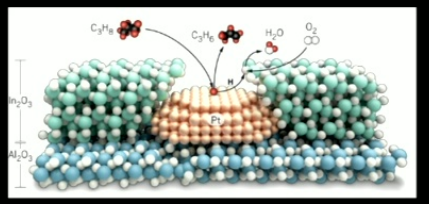


Protein-ligand complexes
(e.g. enzymes)

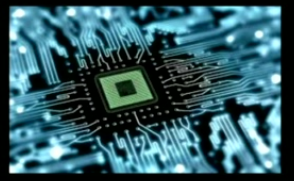
Antigen-antibody binding
(e.g. to HIV, SARS-CoV-2)



Crystal structure
prediction



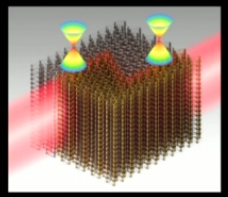
Catalysts



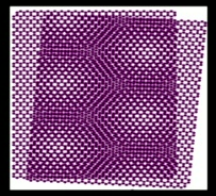
Nanochips

Biochemistry, drug discovery, materials science, nanotechnology

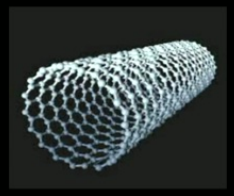
temperature
rconductors



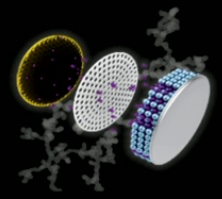
Topological materials



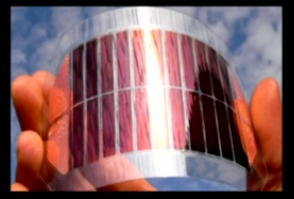
2D materials
(twisted bilayer graphene)



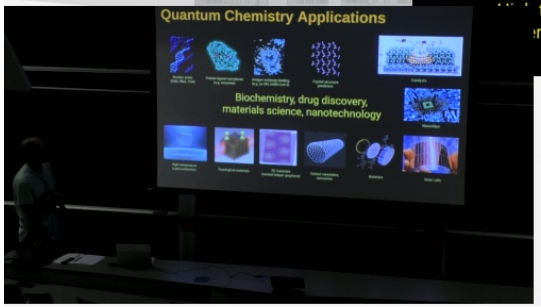
Carbon nanotubes,
nanowires



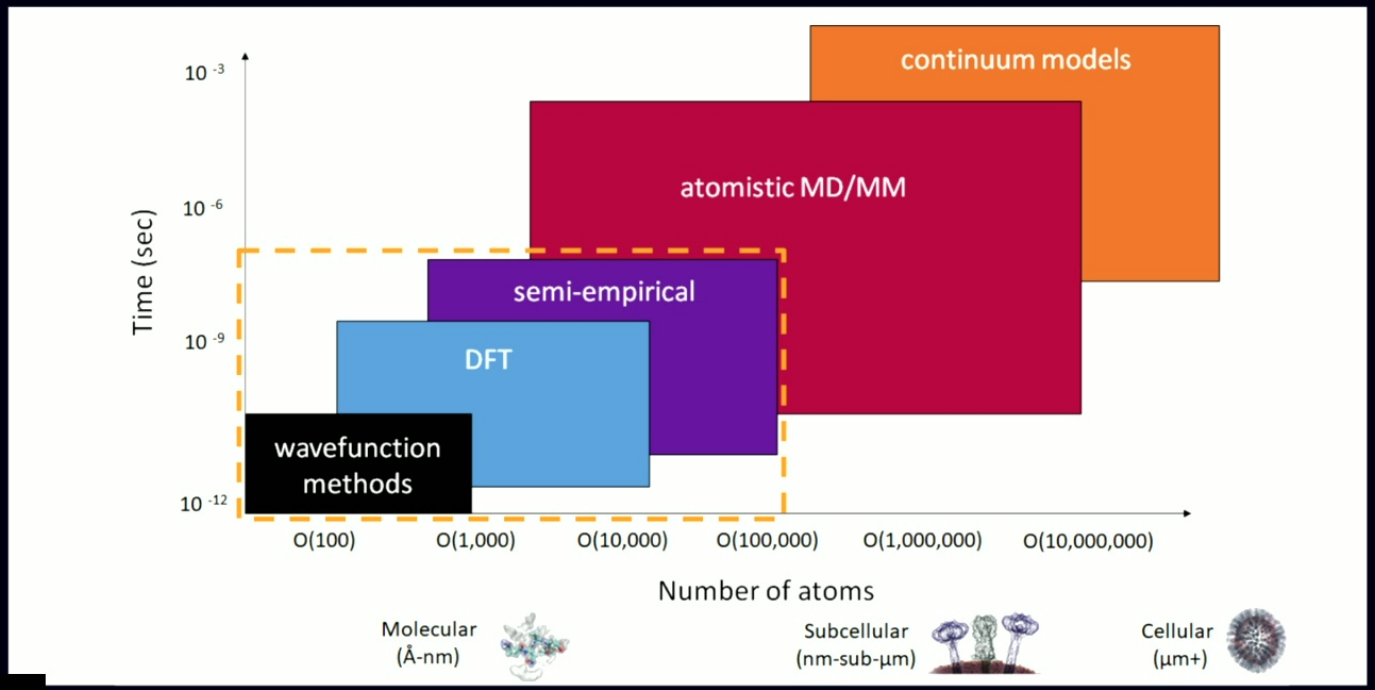
Batteries

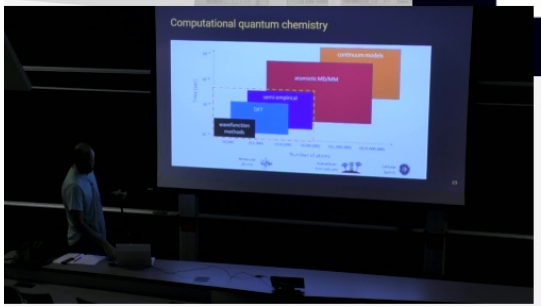


Solar cells



Computational quantum chemistry





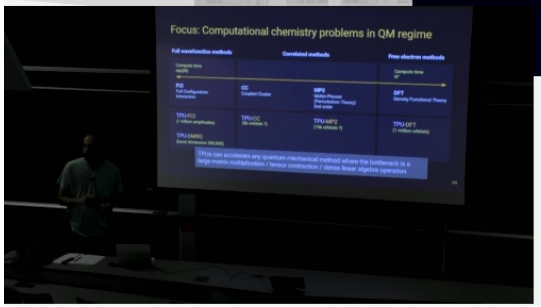
23

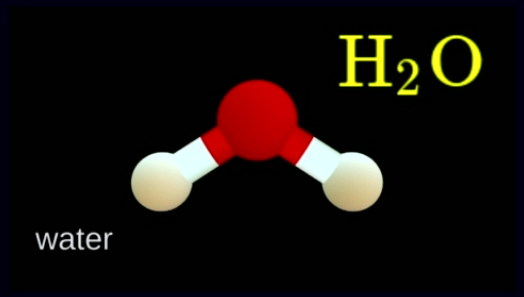
Focus: Computational chemistry problems in QM regime

Full wavefunction methods	Correlated methods		Free electron methods
Compute time exp(N)			Compute time N^3
FCI Full Configuration Interaction	**CC** Coupled Cluster	**MP2** Moller-Plesset (Perturbation Theory) 2nd order	**DFT** Density Functional Theory
TPU-FCI (1 trillion amplitudes)	TPU-CC (5k orbitals ?)	TPU-MP2 (10k orbitals ?)	TPU-DFT (1 million orbitals)
TPU-DMRG (bond dimension 200,000)			

TPUs can accelerate any quantum mechanical method where the bottleneck is a large matrix multiplication / tensor contraction / dense linear algebra operation

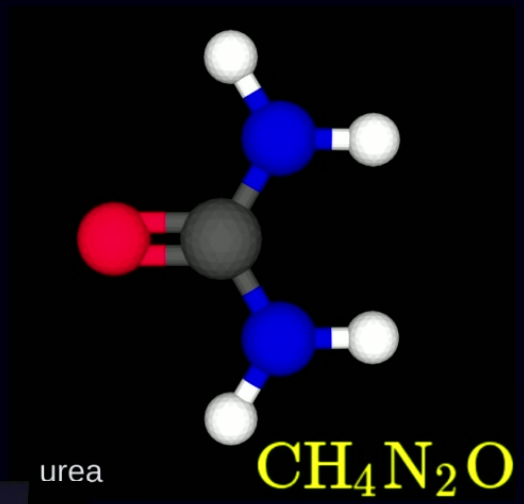
24





$H \rightarrow 1$ electron, 5 orbitals
$O \rightarrow 8$ electrons, 14 orbitals

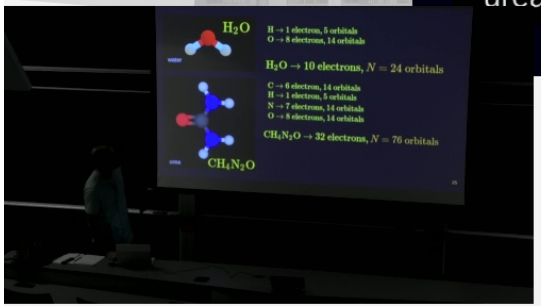
$H_2O \rightarrow 10$ electrons, $N = 24$ orbitals



$C \rightarrow 6$ electron, 14 orbitals
$H \rightarrow 1$ electron, 5 orbitals
$N \rightarrow 7$ electrons, 14 orbitals
$O \rightarrow 8$ electrons, 14 orbitals

$CH_4N_2O \rightarrow 32$ electrons, $N = 76$ orbitals

25



TPU-accelerated Density Functional Theory (DFT)

DFT is a self-consistent field (SCF) method. Each iteration of the DFT loop includes:

H *Hamiltonian* (NxN matrix)

(1) *Hamiltonian* build

$D^{[i-1]} \mapsto H^{[i]}$

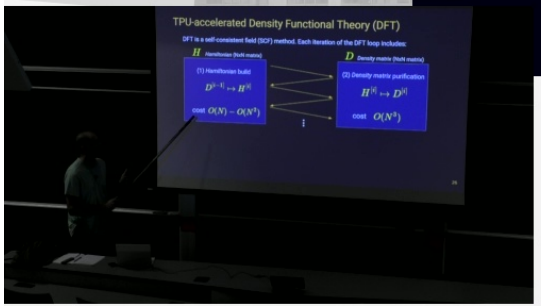
cost $O(N) - O(N^2)$

D *Density matrix* (NxN matrix)

(2) *Density matrix* purification

$H^{[i]} \mapsto D^{[i]}$

cost $O(N^3)$



TPU-accelerated Density Functional Theory (DFT)

DFT is a self-consistent field (SCF) method. Each iteration of the DFT loop includes:

H *Hamiltonian* (NxN matrix)

(1) *Hamiltonian* build

$D^{[i-1]} \mapsto H^{[i]}$

cost $O(N) - O(N^2)$

CPUs

D *Density matrix* (NxN matrix)

(2) *Density matrix* purification

$H^{[i]} \mapsto D^{[i]}$

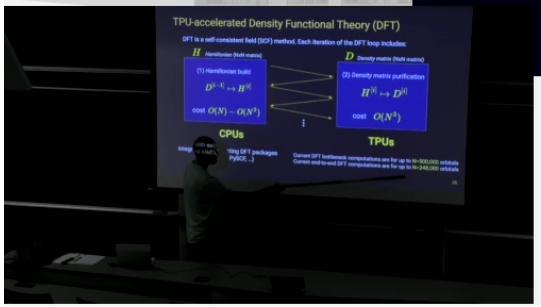
cost $O(N^3)$

TPUs

Integration with existing DFT packages (FHI-AIMS, PySCF, …)

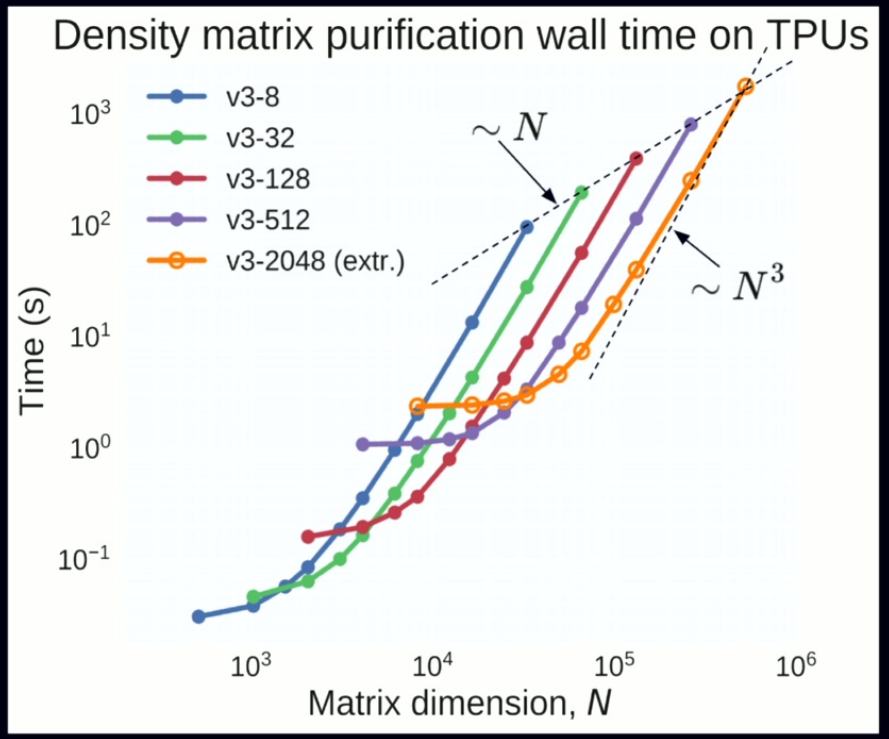
Current DFT bottleneck computations are for up to N=500,000 orbitals
Current end-to-end DFT computations are for up to N=248,000 orbitals

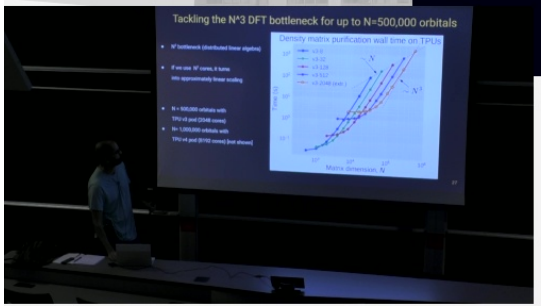
26



Tackling the N^3 DFT bottleneck for up to N=500,000 orbitals

- N^3 bottleneck (distributed linear algebra)
- If we use N^2 cores, it turns into approximately linear scaling
- N = 500,000 orbitals with TPU v3 pod (2048 cores)
- N= 1,000,000 orbitals with TPU v4 pod (8192 cores) [not shown]



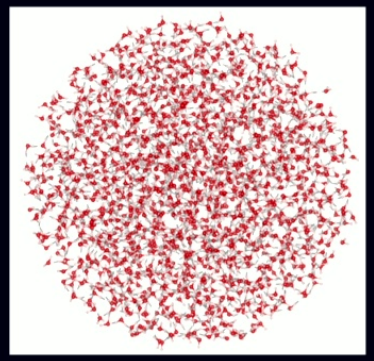


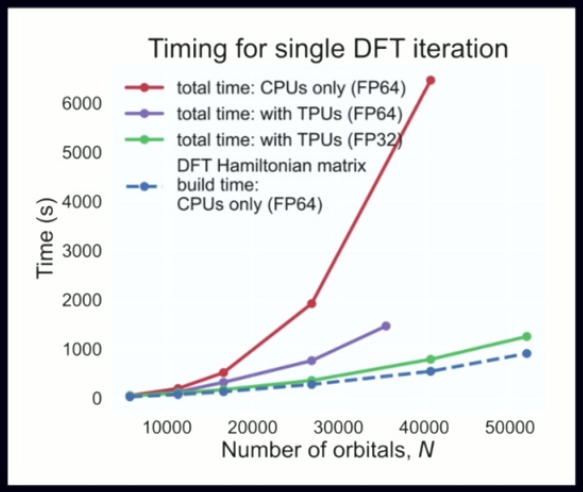
Example of end-to-end DFT computations: large clusters of water molecules

arXiv:2202.01255

Tensor Processing Units as Quantum Chemistry Supercomputers

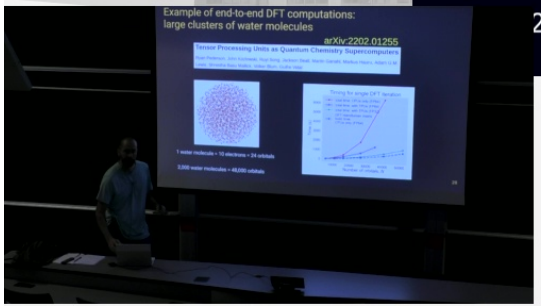
Ryan Pederson, John Kozlowski, Ruyi Song, Jackson Beall, Martin Ganahl, Markus Hauru, Adam G.M. Lewis, Shrestha Basu Mallick, Volker Blum, Guifre Vidal





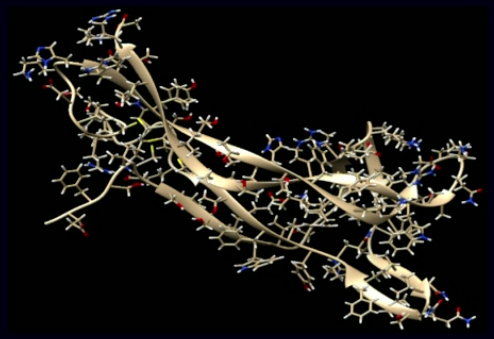
1 water molecule = 10 electrons = 24 orbitals

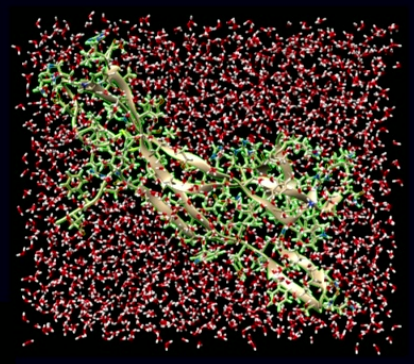
2,000 water molecules = 48,000 orbitals



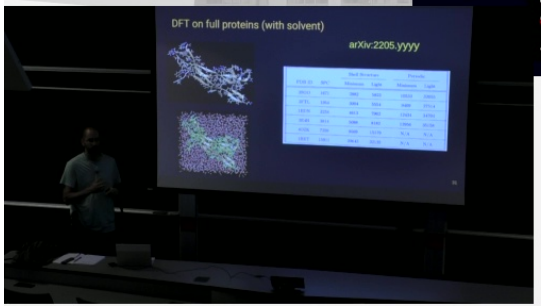
DFT on full proteins (with solvent)

arXiv:2205.yyyy





PDB ID	SPC	Shell Structure		Periodic	
		Minimum	Light	Minimum	Light
3SGO	1671	2882	5823	10533	32055
3FTL	1954	3004	5554	9409	27514
1EDN	3234	4613	7962	12434	34704
3E4H	3814	5088	8182	12956	35158
4OZK	7259	9569	15179	N/A	N/A
1BET	15911	20643	32135	N/A	N/A



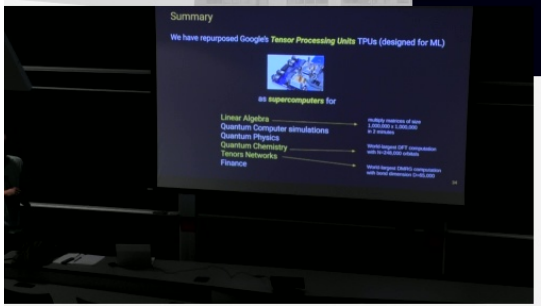
Summary

We have repurposed Google's ***Tensor Processing Units*** TPUs (designed for ML)

as ***supercomputers*** for

Linear Algebra →	multiply matrices of size 1,000,000 x 1,000,000 in 2 minutes
Quantum Computer simulations	
Quantum Physics	
Quantum Chemistry →	World-largest DFT computation with N=248,000 orbitals
Tenors Networks →	World-largest DMRG computation with bond dimension D=65,000
Finance	

34



Series of recent papers describing our work

A. Morningstar, M. Hauru, J. Beall, M. Ganahl, A. GM Lewis, V. Khemani, G. Vidal,
Simulation of quantum many-body dynamics with Tensor Processing Units: Floquet prethermalization, **arXiv:2111.08044** (PRX Quantum)

M. Hauru, A. Morningstar, J. Beall, M. Ganahl, A. Lewis, G. Vidal,
Simulation of quantum physics with Tensor Processing Units: brute-force computation of ground states and time evolution, **arXiv:2111.10466**

A. GM Lewis, J. Beall, M. Ganahl, M. Hauru, S. Basu Mallick, G. Vidal,
Large Scale Distributed Linear Algebra With Tensor Processing Units, **arXiv:2112.09017** (PNSA)

R. Pederson, J. Kozlowski, R. Song, J. Beall, M. Ganahl, M. Hauru, A. GM Lewis, S. Basu Mallick, V. Blum, G. Vidal,
Tensor Processing Units as Quantum Chemistry Supercomputers, **arXiv:2202.01255**

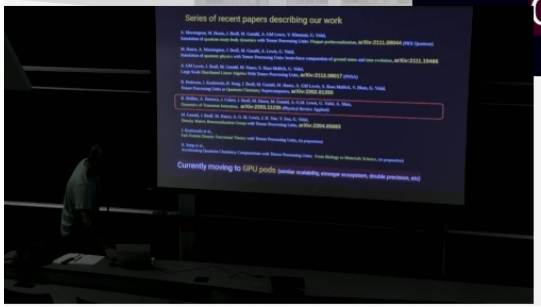
R. Shillito, A. Petrescu, J. Cohen, J. Beall, M. Hauru, M. Ganahl, A. G.M. Lewis, G. Vidal, A. Blais,
Dynamics of Transmon Ionization, **arXiv:2203.11235** (Physical Review Applied)

M. Ganahl, J. Beall, M. Hauru, A. G. M. Lewis, J. H. Yoo, Y. Zou, G. Vidal,
Density Matrix Renormalization Group with Tensor Processing Units, **arXiv:2204.05693**

J. Kozlowski et al.,
Full Protein Density Functional Theory with Tensor Processing Units, (in preparation)

R. Song et al.,
Accelerating Quantum Chemistry Computations with Tensor Processing Units: From Biology to Materials Science, (in preparation)

Currently moving to GPU pods (similar scalability, stronger ecosystem, double precision, etc)

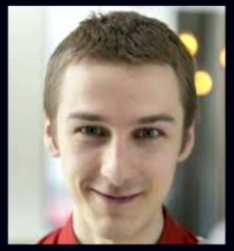


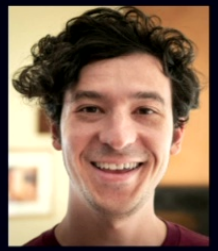
Simulation & Optimization team within Sandbox@Alphabet

(rip Dec 2021)

Martin Ganahl

Research Scientist





THANKS !

Ryan Pederson

Resident Research Scientist

Ruyi Song

Resident Research Scientist

Johnny Kozlowski

Resident Research Scientist

Guifre Vidal

Research Scientist
Research Lead

Shrestha Basu Mallick

Product manager

Matthias Tan

Strategy & Partnerships

Olivier Lacombe

Strategy & Product

Stefan Leichenauer

Eng Lead

Jack Hidary

Director

36