<Cerebras System>
what is the right machine/why is this works
gpus => vision/ showed possibility of value
Ai -> cpu-bad, gpu-less bad

→ new chip for AI acceleration.

Hardware/system/chip design
Ai → calculation trivial/ cpu ← calculation is not trivial / not communication
Communication in memory, core to core is important
new workload → explore, unpack what AI work is

Multiply accumulate, communicate
communicate out of boundary← 1000x power
traditional: tied in one computer/bunch of gpu

build a huge chip → no need to out of boundary, no network
how hard it was?

Radical innovation.
hard and brutal → nobody motherboard, manufacturer,

optimizing the tool for purpose.
ex) F-150 → for kids, bad/ for lumber, good
optimize for big AI work, trillions of parameters.

test ideas not able to test → do work other people could not do.

“Epigenomic language models powered by Cerebras”

<Small Molecule(Drug) screening>
billions data for potential model
supercomputer cluster → bind/ impact

<Argonne National Lab>
Giant workflow → molecular simulation
combining AI with traditional supercomputers.

“Stream-AI-MD:Streaming AI-driven Adaptive Molecular Simulations for Heterogeneous Computing Platforms”
“Intelligent Resolution: Integrating Cryo-EM with AI-driven Multi-resolution Simulations to Observe the Sars-Cov-2 Replication-Transcription Machinery in Action”

Teach the Ai in simulated data

Large Language Model → the same AI for Virus(1000x1000) Genome Sequence

TotalEnergies: Accelerate Multi-Energy Research
"Massively scalable stencil algorithm" 200x faster

Faster than real-time → before finish, get results from CFT model, modifie in real real-world (Digital Twin)

dedicated one thing → gains big enough

chip behaves like a supercomputer: sparsity/ linear algebra

belief we could do it / we are only interested in problem anyone could not solve / afraid of

failure analysis → why failed/ failed in same way or different way

infrastructure builders → package switching => whatsapp
product → fundamental social issue

send back valuable analysis → ex) put sensor to AI chips

Understands what customer works → solves customer's problem

compute/storage/network → problem is permutation of three areas.

storage: how you move data
ex) satellite: data transfer is expensive

grab the useful stuff

What's next step?
we lose alot.