



AUDUN WICKSTRAND IVERSEN
Porteføljeforvalter

EMILIE KRUTNES ENGEN
Porteføljeforvalter

LEO RUNDGREN OLSEN
Junioranalytiker

Hensikten med Disruptive Perspektiver:

Når vi analyserer ulike temaer bruker vi mye tid og mange verktøy (kvartalsrapporter, analyser, dialog med selskapene, bedriftsbesøk, excel, kalkulator og ordmodeller). Ofte lager vi små notater, og noen ganger store notater som vi tenker på som perspektiver. Vi er gamle nok til å vite at det sjeldent finnes sannheter, ofte bare ulike perspektiver.

Disruptive Perspektiver har kun én hensikt: Å dele våre perspektiver på temaer som former vår fremtid. Dette er ikke akademiske notater, innlegg til et leksikon eller anbefalinger om å gjøre noe, kjøpe eller selge noe. Kun god gammeldags informasjonsdeling for å synliggjøre hvordan vi ser på ulike temaer på publiseringstidspunktet. Perspektiver blir ikke mindre, kanskje heller mer, når man deler det. Med det utgangspunktet; ha en fin reise i våre perspektiver.

Innhold

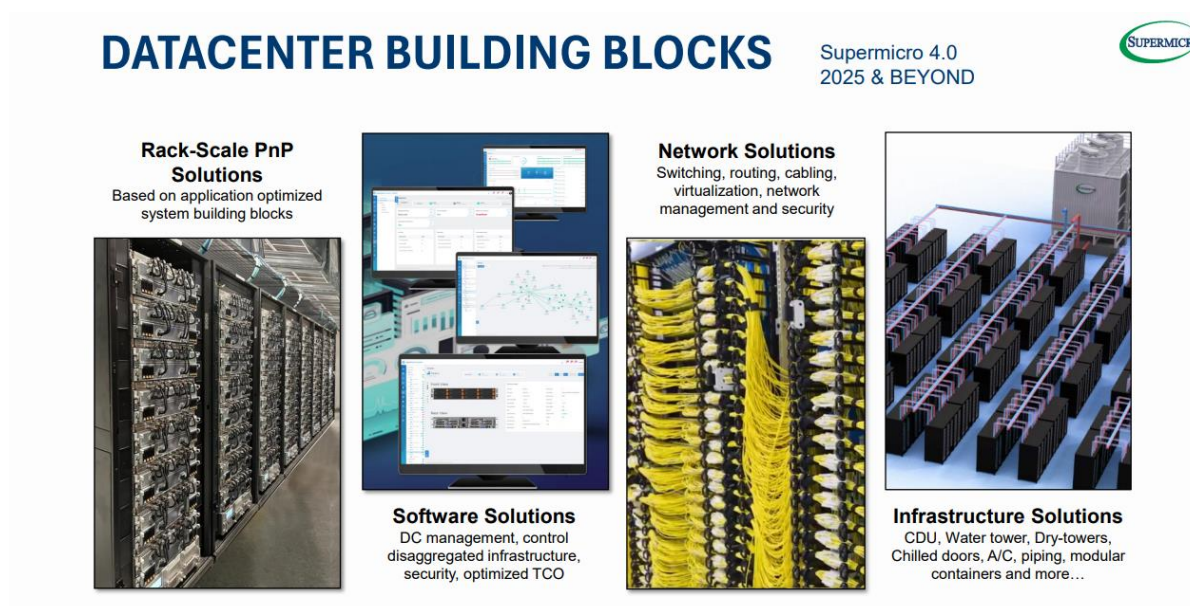
Disclaimer.....	1
Hva er AI-infrastruktur?.....	2
Visjonen	3
Enablers	4
Kategorisering	7
AI-fabrikker.....	10
Hva er en AI-fabrikk?	10
CUDA og Dojo:.....	14
Tokens - Den nye valutaen innen AI	18
Federated Learning	21
Hva er Edge AI?	24
Hva er neuromorfe brikker?	27
Hva er Spiking Neural Networks?.....	31
Tesla og NVIDIA.....	42
Trening: Samspill og selvstendighet.....	43
Handling (Inferens): Edge i bevegelse.....	43
NVIDIAs rolle i trening og handling av kunstig intelligens	46
SNNs og fremveksten av Humanoids.....	49
Fra cloud til edge:	51
Når intelligensen flyttes ut til produktene.....	53
En ny «feedback loop»	53
Hvor finnes de største AI-mulighetene i 2025?.....	54

Disclaimer

Innholdet i denne artikkelen er ikke ment som investeringsråd eller anbefalinger. Har du noen spørsmål om fondene det refereres til, bør du kontakte en finansrådgiver som kjenner deg og din situasjon. Husk også at historisk avkastning i fond aldri er noen garanti for fremtidig avkastning. Fremtidig avkastning vil blant annet avhenge av markedsutvikling, forvalterens dyktighet, fondets risiko, samt kostnader ved kjøp, forvaltning og innløsning. Avkastningen kan også bli negativ som følge av kurstap.

Hva er AI-infrastruktur?

Når vi snakker om kunstig intelligens, er det lett å tenke på apper som ChatGPT, selvkjørende biler eller medisinsk bildediagnostikk. Men sjelden snakker vi om det som gjør alt dette mulig. Selve grunnmuren som Kunstig intelligens (AI) bygger på. Denne grunnmuren kalles AI-infrastruktur, og den består av alt fra maskinvare og programvare til datasentre, nettverk og mennesker som utvikler og drifter teknologien. Uten denne infrastrukturen, får vi ingen AI.



Kilde: Supermicro computer Q1 report

Det som gjør AI-infrastruktur unik, er at den er ekstremt krevende. Der tradisjonell IT-infrastruktur håndterer nettsider eller databaser, krever AI-infrastruktur massiv regnekraft, enorme datamengder, og kontinuerlig utvikling. Moderne språkmodeller trenes på flere hundre milliarder ord og kjøres på tusenvis av GPU'er (Graphics Processing Units) i ukevis.

Og der datainfrastruktur tidligere kunne bygges én gang og brukes i årevis, må AI-infrastruktur være dynamisk. Modellene blir foreldet raskt, nye datasett kommer til, og algoritmer må oppdateres. Derfor er den moderne AI-infrastrukturen ofte skybasert og modulær, med hardware som spesialdesignede GPU-er og TPU-er (Tensor Processing Units), og software som muliggjør kontinuerlig utvikling gjennom MLOps (Maskinlæringsoperasjoner).

Til tross for dette ligner AI-infrastruktur på annen infrastruktur på noen områder: den må være skalerbar, sikker, og avhengig av raske nettverk. Forskjellen ligger i intensiteten og kompleksiteten. AI trenger data. Det er essensielt for videreutvikling, og oppnåelse av AGI (Artificial General Intelligence) og ASI (Artificial Super Intelligence).

Men AI-infrastruktur handler ikke bare om teknologi. Det handler også om hvordan alle disse komponentene henger sammen i et økosystem. Dette økosystemet rommer både tekniske prosesser (som modelltrening og distribusjon), fysisk hardware (som supercomputere og edge-brikker), software-plattformer og konkrete AI-produkter. Alt dette spiller sammen i et system hvor små endringer i én komponent kan få store konsekvenser for helheten. AI-infrastruktur er med andre ord både teknologisk og systemisk, og kanskje den viktigste usynlige motoren bak teknologiske fremskritt de neste tiårene.

Visjonen

Når man bygger en motorvei, handler det ikke bare om asfalt og betong. Det handler om hva man muliggjør. I denne forstand bevegelse, handel, kontakt mellom mennesker. På samme måte handler AI-infrastruktur ikke bare om servere, datasentre og algoritmer. Det handler om å skape en plattform for fremtidens innovasjon.

Det store målet for AI-infrastruktur er å gjøre avansert kunstig intelligens tilgjengelig for alle. Ikke bare for teknologigigantene. Denne demokratiseringen gjør det mulig for startups, forskere, offentlig sektor og enkeltpersoner å bygge løsninger som tidligere var utenkelig. Tenk på klimamodeller, personlig medisin, eller intelligente læringsverktøy. Alt dette avhenger av at infrastrukturen under er tilgjengelig, robust og bærekraftig.

Men for å nå dette målet må vi løse noen fundamentale utfordringer.

- Skala: infrastrukturen må håndtere både enorme datasett og ekstremt komplekse modeller.
- Miljø: dagens AI-trening krever enorme mengder energi, og uten bedre energieffektivitet vil veksten bli uholdbar.
- Tilgjengelighet: AI må kunne brukes av små aktører, ikke bare de med milliardbudsjetter.

- Etikk, sikkerhet og pålitelighet: vi må kunne stole på at systemene vi bygger ikke diskriminerer, feiltolker eller lekker sensitiv informasjon.

Når vi vurderer hvor god AI-infrastrukturen er, må vi derfor se bredt: hvor raskt trenes modellene? Hvor mye energi bruker vi per prediksjon? Hvor tilgjengelig er teknologien for nye aktører? Hvor trygge og rettferdige er systemene i praksis? AI-infrastrukturens suksess kan ikke bare måles i teknisk ytelse. Den må også fungere for samfunnet.

Enablers

Når vi løfter blikket, ser vi at AI-infrastruktur står på skuldrene til 6 kraftige «enablers»:

Først har vi spesialisert hardware (GPU'er) og neuromorfe brikker som gir systemene kraft nok til å lære. Her spiller Moores lov fortsatt en rolle, men den suppleres nå av nye prinsipper som Amdahls lov og innovasjoner innen kvanteberegning.

Deretter kommer data. Uten data, ingen AI. Men det holder ikke med mye data. De må være relevante, rene og kontinuerlig oppdatert. Edge-enheter og syntetiske datasett blir stadig viktigere for å fylle dette behovet.

På software siden finner vi rammeverk som PyTorch og TensorFlow, og MLOps-plattformer som gjør at modeller kan trenes, rulles ut, overvåkes og forbedres - i sanntid.

Cloud-plattformer gir oss elastisk kapasitet og global tilgjengelighet, mens edge computing gir oss lav latens og lokal kontroll. To tilnærminger som utfyller hverandre.

Men teknologien i seg selv er ikke nok. Uten folk som utvikler, vedlikeholder og utfordrer den, kommer vi ikke langt. AI krever tverrfaglig kompetanse fra koding og statistikk til etikk og domeneinnsikt. Her må vi tenke som en lærende organisasjon, hvor kunnskap kontinuerlig oppdateres og deles.

Til slutt er det økonomiske og regulatoriske rammevilkår som setter rammene. Offentlig støtte, skatteincentiver og risikovillig kapital er avgjørende men det er også reguleringer (f.eks GDPR), som sikrer at AI utvikles og brukes ansvarlig.

Hvordan kunne AI-genererte reguleringer sett ut?

En typisk AI-generert regulering ville trolig vært svært detaljert og teknisk. Ved bruk av Natural Language Processing (NLP) og maskinlæringsmodeller ville AI foreslått reguleringer basert på eksisterende regelverk, tekniske spesifikasjoner og offentlige innspill. Et eksempel kunne være:

«AI-systemer klassifisert som høy risiko (definert som en feilrate over 0,1 % i kritiske applikasjoner) skal loggføre alle beslutninger i et revisjonsbart format og undergå årlig bias-testing.»

Slike forslag ville være svært presise, men kunne lett mangle nødvendig fleksibilitet, kulturell forståelse og lokal tilpasning. Elementer som krever menneskelig skjønn.

Den optimale løsningen: komboen mennesker med AI-støtte

Selv om AI ikke bør få fullstendig ansvar for reguleringsarbeidet, betyr ikke det at vi bør ignorere dens potensial. Den beste tilnærmingen er å benytte en hybrid modell, der AI fungerer som et kraftig støtteverktøy.

En kombinasjon av AI og menneskelig vurdering vil gi presise og datadrevne utkast fra AI, som deretter bearbeides av mennesker. Mennesker kan sørge for at reguleringene tar hensyn til etiske, kulturelle og samfunnsmessige dimensjoner. Dette sikrer en balanse mellom teknisk presisjon og praktisk anvendbarhet, samtidig som den demokratiske legitimiteten ivaretas.

Vi kan tydeliggjøre dette med en 2x2-matrise som viser hva slags rolle AI bør spille avhengig av regulerings natur:

Reguleringer

		AI skriver reguleringer alene	Mennesker med AI-støtte
Y-Akse: Reguleringsstil	Fleksibel, Prinsippbasert	+ Rask, konsistent – Generisk	+ Tilpasset, fleksibel – Subjektiv
	Streng, Teknisk	+ Presis, effektiv – Rigid	+ Balansert, praktisk – Mer tid
		X-Akse: AI-involvering	

Disse seks faktorene virker ikke alene, de henger sammen i et dynamisk system. Når en av dem svikter, forplanter det seg gjennom hele økosystemet. Dette er systemteori i praksis: alt henger sammen.

Vi må derfor bygge en infrastruktur som er både kraftig og fleksibel, etisk og effektiv, samt menneskedrevet og maskinforsterket. Og vi må gjøre det på en måte som er åpen for innovasjon, men samtidig bevisst på konsekvensene.

Med dette som utgangspunkt går vi videre til de ulike typene AI-infrastruktur, og hvordan de teknologisk og praktisk er organisert i dagens og morgendagens verden ...

Kategorisering

Når vi nå beveger oss videre inn dybden, dykker vi ned i ulike typer AI-infrastruktur og nøkkelteknologier som utgjør ryggraden i moderne AI. Målet er å gi en forståelig og sammenhengende gjennomgang av hvordan disse teknologiene henger sammen

AI-infrastruktur består av hardware, software, datalagring og nettverk som sammen muliggjør utvikling, trening, drift og vedlikehold av avanserte AI-modeller. Dette helhetlige systemet deler seg naturlig inn i fire hovedområder:

- *Beregningsinfrastruktur*
- *Datainfrastruktur*
- *Software-infrastruktur*
- *Utrullingsinfrastruktur*

Postcards From The Future

Kategorisering av AI-infrastruktur

		Maskinvare	Programvare
Y-Axis: <u>Anvend</u> lighet	Generell	Maskinvare som brukes på tvers av AI og andre applikasjoner, som standard servere, <u>CPU</u> er og generelle lagringsløsninger.	Programvare som brukes i både AI og andre domener, som databehandlingsverktøy (Pandas, Apache Spark) og generelle skyplattformer (AWS, <u>Azure</u>).
	Spesialist	Maskinvare designet spesifikt for AI, som <u>GPU</u> er, <u>TPU</u> er og <u>neuromorfe</u> brikker, optimalisert for parallellprosessering og energieffektivitet.	Programvare utviklet for AI-spesifikke oppgaver, som ML-rammeverk, <u>MLOps</u> -verktøy og AI- <u>API</u> er.
		X-Axis: Underliggende teknologi	

Beregningsinfrastruktur = motoren

Dette er hjertet i AI: spesialtilpassede prosessorer som GPUer (f.eks. NVIDIA H100), er bygd for massiv tallknusing. I tillegg kommer neuromorfe brikker som Intel Loihi, som etterligner hjernens effektivitet, og edge-hardware som NVIDIA Jetson og ARM-enheter som gjør AI kjørbare direkte på enheter som roboter eller sensorer. Større systemer, som NVIDIA DGX-clustere, kombinerer disse for å trene kraftige modeller. Dette følger trendene vi kjenner fra Moores og Amdahls lov: stadig mer kraft og økende behov for parallell ytelse.

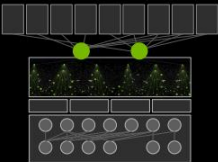
What Is Accelerated Computing?

A full-stack approach: silicon, systems, software

Not just a superfast chip—accelerated computing is a full-stack combination of:

- Chip(s) with specialized processors
- Algorithms in acceleration libraries
- Domain experts to refactor applications

To speed up compute-intensive parts of an application



NVIDIA

Amdahl's law:

The overall system speed-up (S) gained by optimizing a single part of a system by a factor (s) is limited by the proportion of execution time of that part (p).

$$S = \frac{1}{(1 - p) + \frac{p}{s}}$$

For example:

- If 90% of the runtime can be accelerated by 100X, the application is sped up 9X
- If 99% of the runtime can be accelerated by 100X, the application is sped up 50X
- If 80% of the runtime can be accelerated by 500X, or even 1,000X, the application is sped up 5X

Kilde: NVIDIA Investor Relations

Datainfrastruktur er drivstoffet bak intelligensen

Uten gode data nytter ikke regnekraften. Her snakker vi om *data pools* og objektlagring (som Amazon S3), distribuerte databaser og plattformer som Snowflake. Data renses og struktureres gjerne via Spark, Pandas og dedikerte verktøy, og kan komme fra alt fra IoT-sensorer og offentlige åpne datasett, til syntetisk generert data. God datastyring, med anonymisering, kvalitetssikring og GDPR-kompatibilitet, sikrer at GIGO «garbage in, garbage out»-prinsippet unngås. Kort sagt: uten rent drivstoff stopper AI-fabrikken.

Software-infrastrukturen gjør det mulig å bygge komplekse systemer ved å «skjule» det tekniske som ligger under, slik at utviklere kan fokusere på det de skal lage, ikke hvordan alt

fungerer i detalj. Den bruker prinsipper fra abstraksjon, altså å forenkle noe komplisert, slik at flere kan bidra.

Samtidig legger den til rette for innovasjon ved å være tilgjengelig, gjennom bruk av open-source kode. Det betyr at hvem som helst kan se, bruke og forbedre programvaren, noe som gir en felles utviklingsplattform der ideer og løsninger vokser raskere. Rammeverk som TensorFlow, PyTorch og JAX gjør komplisert matematikk tilgjengelig for utviklere. Verktøy som MLflow og Kubeflow håndterer hele livssyklusen for modeller, fra trening til produksjon. Og takket være skytjenester som AWS SageMaker, Azure Machine Learning og Google Vertex AI kan små aktører få tilgang til avansert software uten stor oppstartskostnad. Biblioteker som NumPy og Dask gjør databehandling smidig og rask. Styrken ligger i at mye av dette nettopp er open-source.

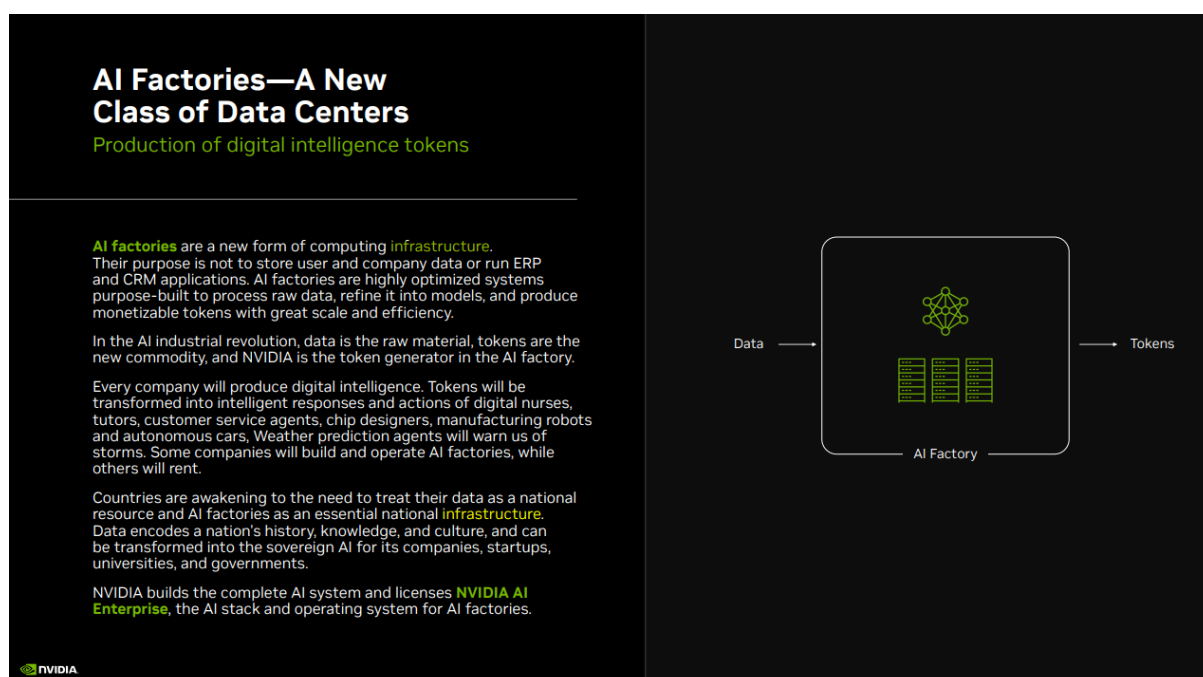
Utrullingsinfrastrukturen bygger på prinsipper fra DevOps (kontinuerlig integrasjon og utrulling), men er tilpasset AI gjennom det som kalles MLOps (machine learning operations). Kort sagt:

DevOps (Development Operations) gjør det lett å oppdatere apper jevnlig. MLOps gjør det mulig å oppdatere AI-modeller jevnlig og sikre at de fortsatt fungerer som de skal, lærer av nye data, og kan styres og overvåkes i drift. Det betyr at AI-modeller ikke bare trenes én gang og ferdig. De vedlikeholdes, forbedres og rulles ut på nytt hele tiden, akkurat som softwareprogrammer. Når en modell er trent, må den fungere i praksis. Her kommer skyplattformer inn, som tilpasser seg belastningen. For mobile anvendelser benyttes teknologier som TensorFlow Lite og ONNX Runtime, slik at AI kan kjøre direkte på enheter. API-er og SDK-er som OpenAI API eller xAI gir enkel integrasjon, mens Prometheus og Grafana sørger for monitorering og kontinuerlig vedlikehold gjennom MLOps-praksiser.

Med denne kunnskapen i bakhodet kan man bevege seg videre for å forstå hva **AI-fabrikker**, **tokens**, **CUDA**, **trening**, **inferens** og **edge-AI** innebærer, og hvordan disse elementene påvirker utviklingen av AI-infrastrukturen i praksis.

AI-fabrikker

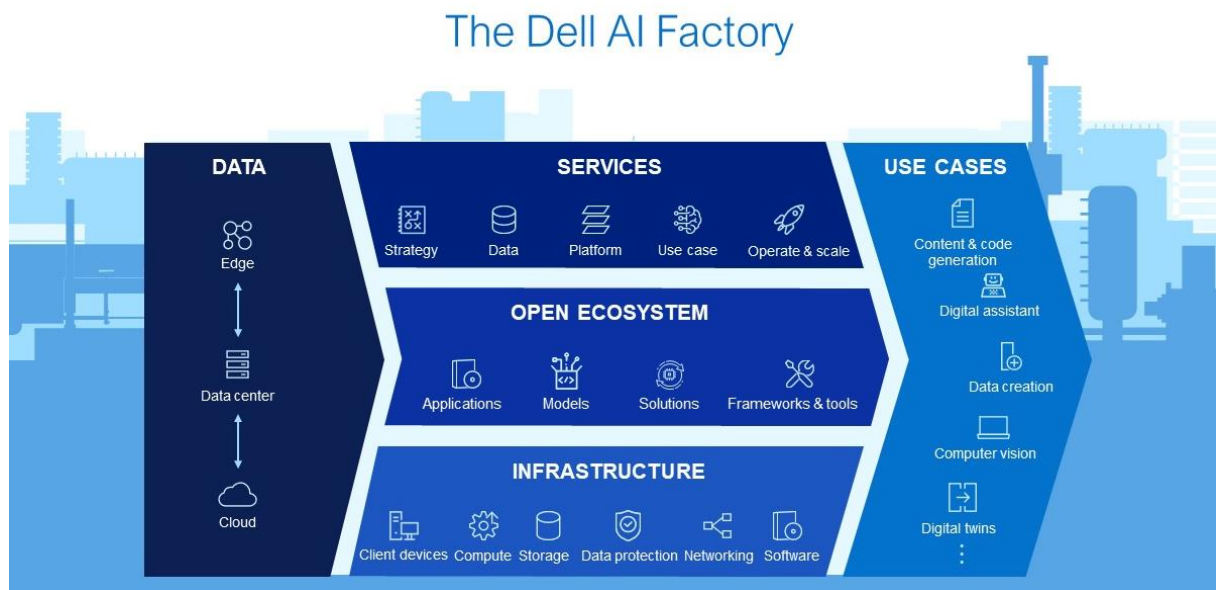
Disse fabrikkene er kjernen i den moderne AI-infrastrukturen. Dette er ikke fabrikker i tradisjonell forstand, men datasentre bygget for én ting: å produsere intelligens. Begrepet ble først popularisert av NVIDIA, og beskriver fasiliteter som kombinerer enorme mengder datakraft med avansert programvare for å trene og drifte AI-modeller i stor skala. I praksis omdanner disse fabrikkene data og energi til modeller som språkmodeller, selvkjørende bilsystemer eller industrielle roboter.



Kilde: NVIDIA IR

Hva er en AI-fabrikk?

En AI-fabrikk er et begrep som beskriver høyt spesialiserte, skalerbare databehandlingsfasiliteter designet for å utvikle, trene, og drifte kunstig intelligens på en industriell skala. Ifølge NVIDIA er AI-fabrikker massive GPU-klynger som fungerer som «fabrikker» for å generere AI-modeller, applikasjoner, og intelligente løsninger, der data og energi er råmaterialer, og AI-tokens (f.eks. prediksjoner, en avgjørelse) er produktet. Disse fasilitetene er ofte skybaserte eller hybridbaserte datasentre som kombinerer kraftig maskinvare, avansert programvare, og datastyring for å støtte AI-arbeidsflyter som dyp læring, naturlig språkbehandling, og autonomi.



Kilde: https://www.dell.com/zh-tw/blog/digital-assistant-solutions-update-with-dell-and-nvidia/?utm_source=chatgpt.com

AI-fabrikker er avgjørende for muliggjørelse av teknologier som selvkjørende biler, humanoids, og AI-agenter, inkludert de som bruker *federated learning* i biler, som vi skal se på senere.

AI-fabrikker er bygget opp av mange ulike teknologier som jobber i lag:

- **Hardware:** Det starter med kraftige brikker som NVIDIA H100 og Google TPU-er, og strekker seg til spesialbrikker som Teslas Dojo D1 og neuromorfe brikker som BrainChip Akida. Disse håndterer enorme mengder data og regneoperasjoner.
- **Software:** Rammeverk som TensorFlow og PyTorch brukes til modellutvikling, mens MLOps-verktøy som MLflow og Kubeflow hjelper med å automatisere prosessene. For optimalisering av ytelse brukes verktøy som NVIDIA TensorRT.
- **Edge:** Når AI kjøres lokalt – som i biler eller humanoide roboter brukes maskinvare som NVIDIA Jetson og programvare som ONNX Runtime.
- **Federated learning:** Modeller kan trenes lokalt på enheter, og oppdateres sentralt – en teknikk særlig kjent fra Tesla.
- **Syntetiske data:** Når ekte data ikke er tilgjengelig, simuleres realistiske datasett ved hjelp av plattformer som NVIDIA Omniverse og CARLA.

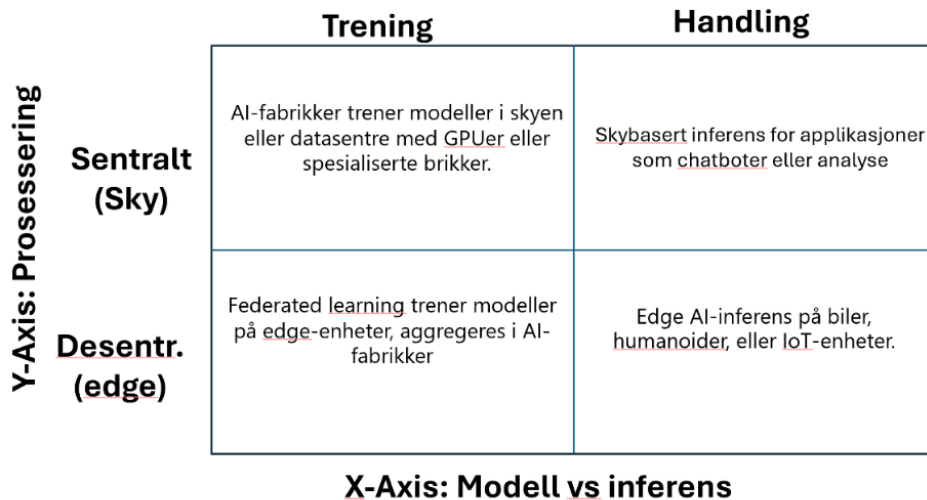
Data er det viktigste råstoffet i en AI-fabrikk. Rådata må samles inn, «vaskes», merkes og analyseres. Verktøy som Apache Spark, Pandas og Databricks brukes til dette. I mange tilfeller analyseres data i sanntid, som i biler eller smarte byer, der strømmer av data sendes inn kontinuerlig. Syntetiske data, laget i simulatorer, brukes også for å supplere eller erstatte virkelige datasett.

Fabrikkene er del av et større nettverk av aktører og industrier som samarbeider:

- **Hardwareleverandører** som NVIDIA, Intel og Tesla bygger brikkene.
- **Skyplattformer** som AWS, Azure og Google Cloud leverer treningsinfrastruktur og API-tjenester.
- **Softwareaktører** som Hugging Face og Databricks lager modeller, rammeverk og analyseverktøy.
- **Bilindustrien**, med Tesla, Waymo, WeRide, Pony.ai m.fl i spissen, bruker AI-fabrikker til å trene autonome kjøresystemer.
- **Robotselskaper** som Boston Dynamics, Agility, 1X, Unitree, Apptronik, Tesla og Figure AI trener humanoids i fabrikkene.
- **Energiselskaper** som Schneider, Emerson, EOSE, ADSE og IREN bidrar med grønn strøm for å gjøre drift bærekraftig.
- **Akademia og forskning** (MIT, Stanford, DeepMind) driver utviklingen av neste generasjons AI.

Vi kan dele AI-fabrikker inn etter hvor de opererer (sky vs. edge) og hva de gjør (trening vs. inferens):

AI-fabrikkøkosystemet



Forklaring av kvadrantene:

1. Trening + sentralisert:

AI-fabrikker trener modeller i skyen eller datasentre med GPUer eller spesialiserte brikker. Eksempel: NVIDIA DGX for GPT-4, Tesla Dojo for FSD.

Aktører: NVIDIA, Tesla, Google, AWS.

2. Handling + sentralisert:

Skybasert inferens for applikasjoner som chatboter eller analyse. Eksempel: Azure AI for inferens, NVIDIA Triton.

Aktører: Microsoft, Google, NVIDIA.

3. Trening + desentralisert:

Federated learning trener modeller på edge-enheter, aggregeres i AI-fabrikker. Eksempel: Tesla-biler trener FSD lokalt.

Aktører: Tesla, Waymo, Google.

4. Handling + desentralisert:

Beskrivelse: Edge AI-inferens på biler, humanoider, eller IoT-enheter. Eksempel: NVIDIA DRIVE Orin, BrainChip Akida.

Aktører: NVIDIA, Tesla, Intel, BrainChip.

AI-fabrikker er økosystemer. Hver komponent, fra maskinvare og programvare til datasikkerhet og energikilde, spiller en rolle. Systemteori hjelper oss å forstå hvordan disse

delene henger sammen. Innovasjonsteori forklarer hvorfor aktører som NVIDIA og Tesla lykkes i å disrupte hele bransjer.

CUDA og Dojo:

To av de mest sentrale teknologiene i dagens AI-infrastruktur er CUDA og Dojo. De representerer to ulike strategier for å løse samme utfordring: Hvordan trene og bruke avanserte AI-modeller effektivt. Mens CUDA er bransjestandarden utviklet av NVIDIA og brukes bredt på tvers av industrier, er Dojo Teslas egenutviklede løsning, spesialisert for deres videobaserte AI-modeller som brukes i selvkjørende biler og humanoide roboter. Sammen illustrerer de både samarbeid og konkurranse i utviklingen av AI-fabrikker.

Hva er CUDA?

CUDA (Compute Unified Device Architecture) er NVIDIAs plattform for generell databehandling på GPUer. I praksis betyr dette at utviklere kan bruke NVIDIAs grafikkprosessorer til å utføre krevende beregninger som tidligere måtte gjøres av CPU'er. Dette gjør CUDA ideelt for AI-trening, der store datamengder og komplekse beregninger må håndteres parallelt.

CUDA er dypt integrert i AI-økosystemet gjennom støtte for populære rammeverk som TensorFlow og PyTorch (SNNs), og tilbyr optimaliserte biblioteker som cuDNN (for deep learning) og cuBLAS (for lineær algebra). I kraftige systemer som NVIDIA DGX og H100 GPUer, med titusener av kjerner, akselererer CUDA trening av store språkmodeller, nevralt nettverk og simuleringer.

Why Accelerated Computing?

Advancing computing in the post-Moore's law era

Accelerated computing is needed to tackle the most impactful opportunities of our time—like AI, climate simulation, drug discovery, ray tracing, and robotics.

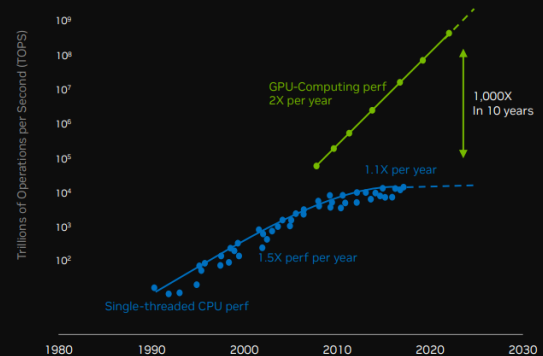
NVIDIA is uniquely dedicated to accelerated computing—working top-to-bottom, refactoring applications and creating new algorithms, and bottom-to-top inventing new specialized processors, like RT Cores and Tensor Cores.

"It's the end of Moore's law as we know it."

—John Hennessy, Oct 2018

"Moore's law is dead."

—Jensen Huang, GTC 2013

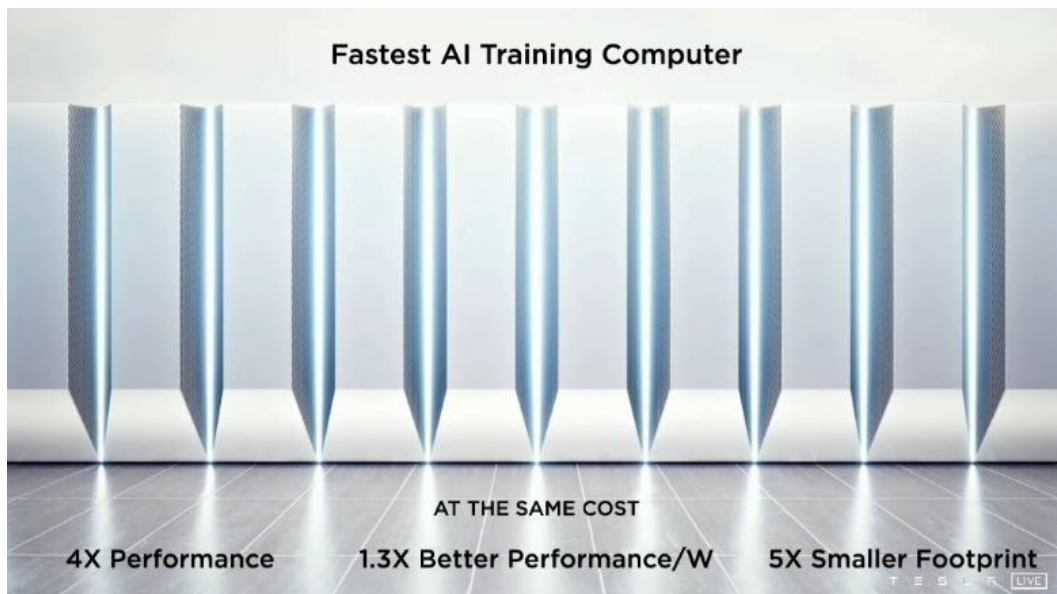


Kilde: NVIDIA IR

I inferens, altså selve bruken av modellene etter trening, muliggjør CUDA rask respons i både sky og edge-enheter. Med verktøy som TensorRT kan for eksempel Teslas selvkjørende biler gjøre bildegjenkjenning på under 10 millisekunder. Dette er avgjørende i applikasjoner der hvert millisekund teller.

Hva er Dojo?

Dette er Teslas spesialutviklede AI-supercomputer, bygget for én hovedoppgave: å trene deres egne modeller for autonom kjøring og robotikk. I motsetning til CUDA, som er bredt anvendelig, er Dojo skreddersydd for Teslas videobaserte arbeidsflyt og store mengder bilflåtedata.



Kilde: Tesla AI day 2022

Kjernen i Dojo er D1-brikken, som er utviklet av Tesla og integrert i store «wafers» for maksimal båndbredde og lav latenstid. En enkelt wafer kan levere opptil 9 petaFLOPS og håndtere petabyte med videoopptak. Perfekte forhold for å trene komplekse modeller som skal tolke trafikk, forutse hendelser og planlegge kjøremønstre.

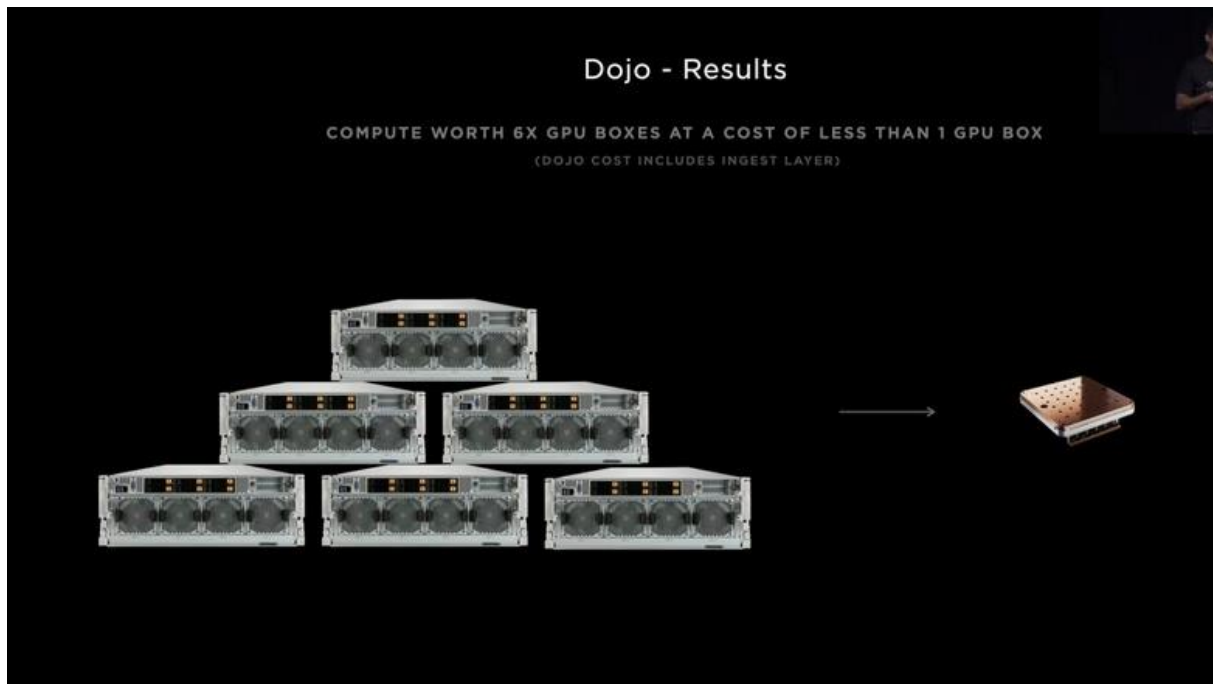
Dojo er tett integrert i Teslas «federated learning»-system. Her samles innsikt fra millioner av biler lokalt og sendes tilbake til Dojo, som oppdaterer modellene uten at rådata forlater bilene. På denne måten forbedres systemet kontinuerlig, samtidig som personvernet ivaretas.

Samspill

CUDA og Dojo brukes ofte i kombinasjon i Tesla-universet. I dag brukes CUDA fortsatt til mye av treningen og nesten all inferens, spesielt på Teslas edge-enheter som HW3 og HW4. Samtidig utvikles Dojo for å ta over stadig mer av treningsarbeidet, særlig for videobasert analyse, og på sikt redusere avhengigheten av NVIDIA.

Dette illustrerer en interessant dynamikk: Mens Tesla samarbeider med NVIDIA og benytter deres teknologi, bygger de samtidig en spesialisert løsning som kan bli en konkurrent på lengre sikt.

I desentraliserte AI-arbeidsflyter som *federated learning*, spiller både CUDA og Dojo viktige roller. CUDA akselererer lokal trening i bilene, mens Dojo fungerer som en sentral node som samler og prosesserer oppdateringene. Denne kombinasjonen gjør det mulig å trene modeller på en effektiv, energieffektiv og sikker måte. Ideelt for storskala autonomi.



Kilde: Tesla AI day 2022

CUDA og Dojo sees som ulike subsystemer i det overordnede AI-økosystemet: CUDA som en generell løsning med bred støtte, og Dojo som en spesialisert motor for Teslas vertikalt integrerte infrastruktur. Innovasjonsteori peker på hvordan CUDA i sin tid revolusjonerte AI ved å gjøre GPU-beregninger tilgjengelig, mens Dojo representerer en ny bølge av spesialiserte, bedriftsinterne løsninger.

Bærekraftsperspektivet er også relevant: Dojo er designet for å være mer energieffektiv enn tradisjonelle GPU-klynger, og støtter Teslas mål om å kutte kostnader og karbonavtrykk. Samtidig spiller personvernteori en rolle, spesielt når flåtedata holdes lokalt og kun modelloppdateringer deles.

Vi kan oppsummere forskjellene og rollene slik:

CUDA og Dojo i trening og handling

		Treningsmodell	Handling
Y-Axis: Anvendelse	Generell	CUDA akselererer trening på tvers av industrier med NVIDIA GPUer.	CUDA støtter inferens i skyen og på edge via TensorRT og DRIVE Orin.
	Nisje	Dojo trener videobaserte modeller for Tesla FSD og Optimus..	Dojo har begrenset rolle i inferens, primært for skybasert testing..
		X-Axis: Modell eller inferens	

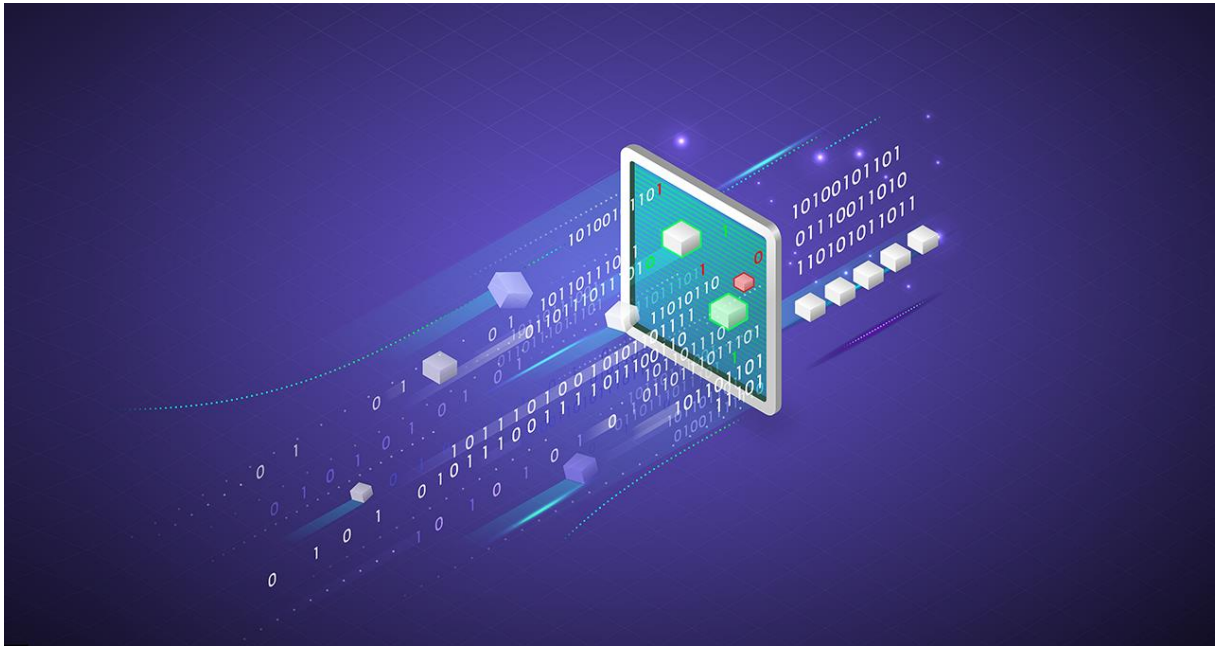
I sum representerer CUDA og Dojo to ulike tilnærminger til AI-infrastruktur. Den ene er universell og åpen for alle, den andre er smal og målrettet. Men begge er avgjørende for å forstå hvordan AI faktisk produseres, og hvordan fremtidens autonome systemer utvikles i praksis.

Tokens - Den nye valutaen innen AI

Han med skinnjakka Jensen Huang i NVIDIA, posisjonerer selskapet som en hjørnestein i en ny industriell revolusjon. AI-fabrikker er datadrevne systemer bygget på NVIDIAs GPU'er og AI-plattformer som Blackwell. Disse fabrikkene produserer tokens som er en ny enhet for økonomisk verdi. Elektrisitet muliggjorde industriell produksjon, og internett drev informasjonsøkonomien. Huang ser AI-infrastruktur som neste bølge, der intelligens, målt som tokens, blir en universell ressurs som transformerer økonomier. Vi starter rolig og ender opp i en 2x2 matrise.

Huang sier NVIDIA ikke lenger bare er et chip-selskap, men en leverandør av infrastruktur for AI-drevne økonomier. Dette inkluderer GPU'er, AI-plattformer, og software som CUDA og Blackwell som driver alt fra selvkjørende biler til humanoide roboter. NVIDIA har tradisjonelt kun solgt GPU'er Disse har vært særlig brukt i gaming, kraftige pc'er og har så langt vært byggesteinene i de nye datasentrene.

AI-fabrikkene, som vi var inne på, prosesserer enorme mengder informasjon for å trene og kjøre AI-modeller. I stedet for å produsere klær eller brød, produserer disse "fabrikkene" tokens. En ny måleenhet for AI-output som tekst, bilder, beslutninger eller handlinger generert av modeller. Én token kan være et ord i en tekst, en piksel i et bilde, eller en beslutning. Modeller som ChatGPT genererer tekst-tokens, mens programvaren i Teslas robottaxier produserer beslutningstokens i sanntid.



Kilde: [*Explaining Tokens — the Language and Currency of AI | NVIDIA Blog*](#)

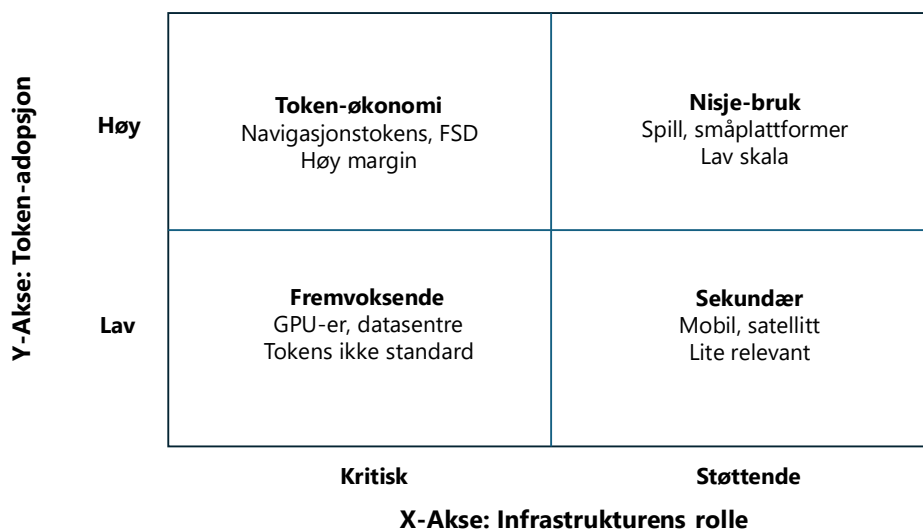
Huang foreslår at tokens vil bli en ny valuta for produktivitet, lik hvordan kilowatt-timer måler elektrisitet eller gigabyte måler datatrafikk. For eksempel kan en AI-agent som styrer en robotaxi produsere «beslutningstokens» per time. Huang har også en visjon om å måle økonomisk output i «tokens per time» antyder en verden der AI-produktivitet kvantifiseres direkte, lik hvordan vi måler fabrikkproduksjon i enheter per time. For eksempel kan en humanoid robot produsere 1,000 beslutningstokens per time i en fabrikk eller 500 emosjonelle tokens i eldreomsorg. Det å etablere en generell måleenhet i AI blir nok vanskeligere når man beveger seg fra ordmodellene og over i andre domener. Å utvide tokens til for eksempel beslutningstokens for FSD og handlingstokens for roboter er logisk, men krever standardisering. For eksempel kan en robotaxi produsere "navigasjonstokens" per kilometer.

Hvis tokens blir en måleenhet, kan de prises direkte som f.eks \$0.01 per token for tekstgenerering eller \$0.10 per beslutningstoken for selvkjøring. Forretningsmodeller rundt tokens vil kunne skape høymargin-inntekter, lik Apples App Store.

I vår 2x2-matrisen kaller vi X-aksen for infrastrukturens rolle, og skiller mellom kritisk og støttende) og på Y-aksen har vi grad av token-adopsjon fra høy til lav:

Postcards From The Future

AI Tokens



Den første kvadranten kaller vi «Fremvoksende AI-infrastruktur» og er der kritisk og lav token-adopsjon møtes. NVIDIAs AI-fabrikker er kritiske for selvkjørende biler, roboter, og AI-agenter til ordmodellene, men tokens er ikke en standard måleenhet.

Den andre kvadranten kaller vi «Token-drevet økonomi» og finner vi der kritisk møter høy token-adopsjon

Hvis tokens blir standard, som for eksempel navigasjonstokens for FSD, emosjonelle tokens for roboter, vil NVIDIA bli en av de dominerende i en token-basert økonomi. Marginer forblir høye pga. software-dominans.

Den tredje kvadranten kaller vi «Sekundær infrastruktur» og finner vi der støttende og lav token-adopsjon møtes.

Teknologier som satellitter, mobiltelefoner og droner støtter AI-fabrikker, men er ikke kritiske. Tokens er irrelevante her.

I den fjerde kvadranten finner vi «Nisje token-bruk». Her er infrastrukturen støttende, men med høy token-adopsjon

Små aktører kan bruke tokens i nisjer, men mangler skala. Marginer er lave pga. konkurranse. Det blir som en litt mislykka krypto valuta eller spill.

Om han med skinnjakka sine visjoner om en ny industriell revolusjon, med AI-fabrikker som produserer tokens, blir en realitet vet vi ikke. Men vi følger med, og det bør du også gjøre.

Federated Learning

Et av de mest spennende gjennombruddene innen AI-infrastruktur de siste årene er fremveksten av *Federated Learning* (FL). Dette er en ny måte å trene AI-modeller på. Ikke i et sentralisert datasenter, men direkte på desentraliserte enheter som smarttelefoner, roboter og sensorer.

I stedet for å samle inn store mengder rådata i skyen, blir modellen sendt ut til hver enkelt enhet. Der trenes den lokalt på brukerens data, og kun små oppdateringer, som modellens vektorer, sendes tilbake. Disse blir så kombinert på en sentral server for å forbedre den globale modellen. På denne måten forblir data der de hører hjemme: på enheten. Det gir bedre personvern, lavere båndbreddebruk, og mulighet for sanntidsforbedring på enheten.

Hvordan det fungerer - steg for steg:

1. **Modellen distribueres:** En sentral server sender ut en startmodell til tusenvis av enheter.
2. **Lokal trening:** Hver enhet trener modellen på egne data.
3. **Oppdateringer sendes tilbake:** Kun modelloppdateringer (ikke data) returneres.
4. **Global aggregering:** En sentral server kombinerer oppdateringene – ofte med en metode kalt *Federated Averaging*.

5. **Ny runde:** Den forbedrede modellen sendes ut igjen – og prosessen gjentas.

Hvorfor er det så viktig? Jo, Federated learning har flere unike fordeler:

- **Personvern:** Data forlater aldri enheten – i tråd med reguleringer som GDPR.
- **Effektivitet:** Mindre behov for tung datalagring og skytilgang.
- **Skalerbarhet:** Mulighet for læring på milliarder av enheter.
- **Sanntidslæring:** Modellene kan forbedres kontinuerlig, basert på faktisk bruk.

Dette gjør teknologien særlig nyttig i *Edge AI*. Altså når maskinlæring skjer nær brukeren, for eksempel i en selvkjørende bil eller en robot i helsevesenet.

Eksempler

- **Google** bruker FL i Gboard-tastaturet, slik at skrivemønstre forbedrer modellen uten at noe sendes til skyen.
- **Apple** benytter det i Siri og QuickType.
- **Tesla** kombinerer det med Dojo for å forbedre FSD-modellen basert på data fra bilflåten.

På hardware siden er neuromorfe brikker som Intel Loihi og BrainChip Akida skreddersydd for slike scenarier. De er designet for å lære på en strømeffektiv måte, og fungerer godt på edge-enheter der regnekraft og batteri er begrenset.

Tradisjonell AI-trening foregår sentralt i store datasentre drevet av GPU-er og massive datasett. Federated learning snur dette på hodet. Den desentraliserer ikke bare læringen, men kombinerer den med handling. Det vil si at enheten både bruker modellen (inferens) og forbedrer den (trening).

Vi kan tenke oss fire typer AI-arbeidsflyter i en 2x2 matrise:

Federated Learning i AI-infrastruktur

		Trening	Handling
Y-Axis: Prosessering	Sentralt	Tradisjonell modelltrening i skyen eller datasentre med GPUer/TPUer.	Inferens i skyen for skalerbare applikasjoner som chatboter eller anbefalingssystemer
	Desentr. (Edge)	Federated learning, der modeller trenes lokalt på edge-enheter og aggregeres sentralt.	Edge AI-inferens med neuromorfe brikker eller optimaliserte prosessorer.
		X-Axis: Modelltrening eller inferens	

Federated Learning plasserer seg i nedre venstre hjørne: desentralisert trening, nær brukeren.

Selv om teknologien har stor oppside, finnes det også svakheter:

- **Varierende datakvalitet:** Heterogene data på tvers av enheter gir utfordringer i aggregert modell.
- **Begrenset maskinvare:** Ikke alle enheter kan trene store modeller effektivt.
- **Kommunikasjon:** Modelloppdateringer må fortsatt sendes over nett – særlig utfordrende i 5G-frie områder.
- **Sikkerhet:** Selv uten rådata kan man i teorien rekonstruere sensitiv informasjon fra modelloppdateringer. Derfor brukes ofte teknikker som *differensielt personvern*.

Federated learning representerer en overgang til *distribuert AI-infrastruktur*, støttet av teorier fra systemtenkning, personvern og bærekraft. Google introduserte begrepet i 2016, og siden har teknologien blitt drevet frem av behov for datasikkerhet og skalerbarhet i en hyperkoblet verden.

I en fremtid der milliarder av enheter er koblet til nett, og der data produseres overalt, fremstår *Federated Learning* som en logisk vei videre – både for personvern, effektivitet og bærekraft.

Hva er Edge AI?

Edge handler om å bringe intelligensen nærmere kilden. I stedet for at data sendes til store datasentre for prosessering, kjører kunstig intelligens direkte på enheter ute i felt, såkalte *edge-enheter*. Det kan være alt fra smarttelefoner og droner til sensorer i smarte byer eller roboter på en fabrikk. Disse enhetene bruker forhåndstreinte AI-modeller til å ta avgjørelser lokalt (ofte i sanntid) uten at data må sendes til skyen. Resultatet er lavere forsinkelser, bedre personvern, og lavere energiforbruk.

Hvorfor er dette viktig?

- **Sanntidsbeslutninger:** Essensielt for autonome kjøretøy og smarte kameraer som må reagere på millisekunder.
- **Personvern:** Ved å beholde data på enheten reduseres risikoen for datalekkasjer og man møter krav fra GDPR og lignende reguleringer.
- **Energieffektivitet:** Ved å unngå tunge overføringer til skyen reduseres strømforbruket.
- **Skalerbarhet:** Gjør det mulig å håndtere milliarder av IoT-enheter samtidig – uten å overbelaste sentraliserte systemer.

Edge AI støttes av spesialisert maskinvare og programvare. Neuromorfe brikker som Intel Loihi, BrainChip Akida og IBM TrueNorth etterligner hjernen og kjører AI ekstremt energieffektivt. Samtidig finnes optimaliserte programvareverktøy som TensorFlow Lite, ONNX Runtime og Lava, utviklet for å få mest mulig ut av lavressursenheter.

- **Intel Loihi:** Forskningstung neuromorf brikke for robotikk og sanntidssystemer
- **BrainChip Akida:** Kommersielt rettet og klar for bruk i IoT-enheter
- **IBM TrueNorth:** Banebrytende innen neuromorf forskning, men mindre brukt kommersielt

Historisk utvikling

Kunstig intelligens i edge er et resultat av en langsom, men tydelig teknologisk forskyvning:

- **1950–1980:** Sentraliserte stormaskiner – ingen Edge AI
- **1980–2000:** Nevrale nettverk vokser frem, men maskinvaren er for svak
- **2000–2015:** GPU-revolusjonen muliggjør kraftig skybasert AI
- **2015–2020:** Tidlige rammeverk som TensorFlow Lite gjør AI på mobilen mulig
- **2020–2025:** Edge AI vokser kraftig, drevet av 5G, IoT og energieffektive brikker

Fordeler:

- Svært lav latenstid (f.eks. <10 ms)
- Minimalt energiforbruk (milliwatt-nivå)
- Sterkt personvern
- Tåler massiv skala

Begrensninger:

- Begrenset regnekraft på enhetene
- Vanskeligere å utvikle og optimalisere modeller
- Mindre modne verktøy og økosystem
- Foreløpig begrenset til spesifikke applikasjoner

For å plassere Edge'n i det store bildet, kan vi bruke en 2x2-matrise:

[Postcards From The Future](#)

Edge AI i AI-infrastruktur

		Generell	Spesialisert
Y-Axis: Bruksområdet	Sentralt	Skybaserte systemer som bruker GPUer/TPUer for generelle AI-oppgaver.	Neuromorfe systemer i forskningsmiljøer for å utvikle SNNs og hjernelignende AI.
	Desentr.	Generelle prosessorer for edge-enheter, brukt for både AI og andre oppgaver	Edge AI drevet av neuromorfe brikker som Loihi, Akida og TrueNorth for lavenergi, sanntidsprosessering.
		X-Axis: Teknologiens anvendelighet	

Forklaring av kvadrantene

1. Generell og sentralisert:

Skybaserte systemer som bruker GPUer/TPUer for generelle AI-oppgaver. Eksempler: NVIDIA DGX, Google TPU v4.

Betydning: Dominerer storskala modelltrening, men energikrevende.

2. Spesialisert og sentralisert:

Neuromorfe systemer i forskningsmiljøer for å utvikle SNNs og hjernelignende AI. Eksempler: Intel Hala Point, IBM TrueNorth i forskning.

Betydning: Viktig for teoretisk fremgang, men mindre kommersialisert.

3. Generell og desentralisert:

Generelle prosessorer for edge-enheter, brukt for både AI og andre oppgaver. Eks: Qualcomm Snapdragon, ARM Cortex.

Betydning: Fleksible, men mindre effektive for AI-spesifikke oppgaver.

4. Spesialisert og desentralisert:

Edge AI drevet av neuromorfe brikker som Loihi, Akida og TrueNorth for lavenergi, sanntidsprosessering. Eksempler: Akida i IoT-sensorer, Loihi i robotikk.

Betydning: Driver bærekraft og sanntids-AI, men begrenset til nisjeapplikasjoner

Denne matrisen viser at Edge i dag hovedsakelig befinner seg i kategorien *spesialisert og desentralisert*. Den retter seg mot nisjeapplikasjoner, men potensialet er enormt, særlig når det gjelder bærekraft og sanntidssystemer.

Fremover vil Edge AI spille en stadig viktigere rolle og teknologier som Akida 2 og Loihi 2 vil løfte ytelsen. Samtidig vil 6G og IoT åpne dørene for nye applikasjoner i alt fra helse og industri til robotikk og smarte byer. En endring i hvordan vi tenker om intelligens i systemer: Fra sentral kontroll til lokal autonomi.

Hva med fremover?

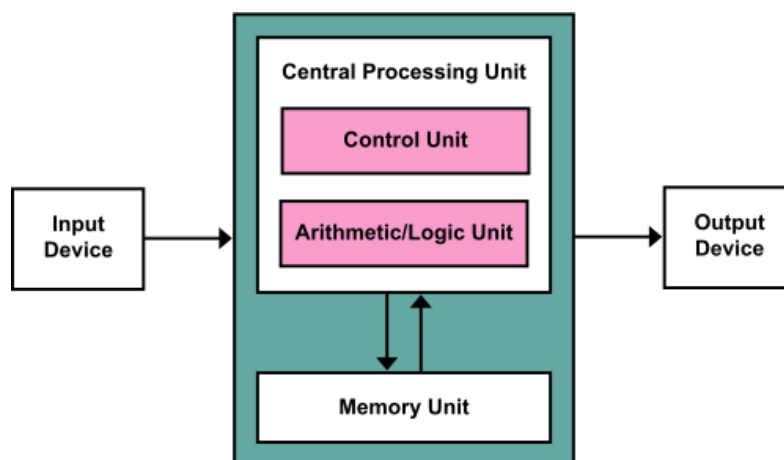
Edge AI vil vokse med nye versjoner av Akida og Loihi, og med 6G + IoT kan helse-, industri, robot- og smarte by-tjenester ta helt nye former. Vi ser et paradigmeskifte: fra sentral intelligens til lokal autonomi der AI ikke bare kjøres, men også lærer og tilpasser seg løpende uten eksponering.

AI-infrastruktur er et komplekst, hierarkisk system fra fysisk hardware til global distribusjon og monitorering. Det inkluderer kraftige AI-fabrikker, tokens, desentralisert læring og real-time intelligens på kanten.

Hva er neuromorfe brikker?

Neuromorfe brikker er en ny type mikroprosessorer som er inspirert av hjernens måte å behandle informasjon på. I stedet for å bruke den tradisjonelle datamaskinarkitekturen, der prosessering og minne er atskilt, etterligner neuromorfe brikker hvordan nevroner og synapser jobber sammen i hjernen. Resultatet er en distribuert, parallell og svært energieffektiv måte å utføre beregninger på. Skreddersydd for oppgaver som mønstergjenkjenning, sensorisk prosessering og AI-inferens med lavt strømforbruk.

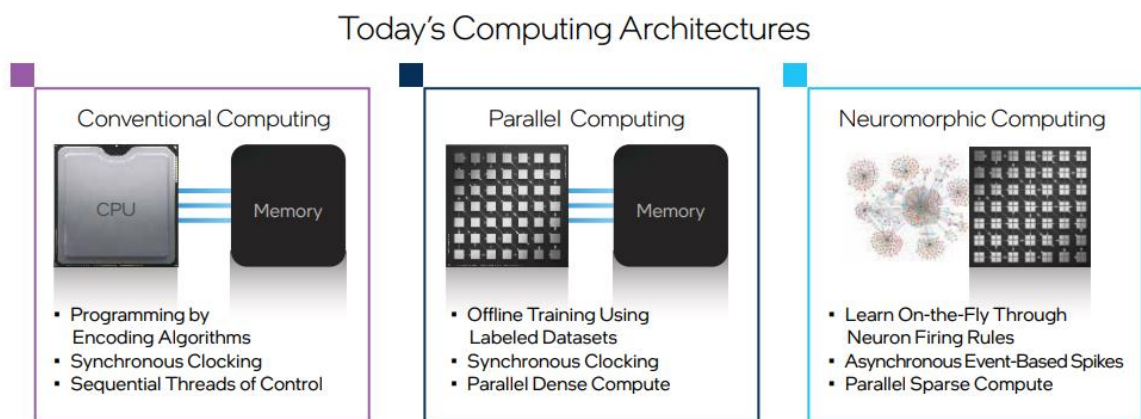
Brikkene bruker såkalte spiking neural networks (SNN), der "nevroner" fyrer elektriske signaler («spikes») kun når visse terskler er nådd. Akkurat som i den biologiske hjernen. Denne hendelsesdrevne prosesseringen gjør dem langt mer energieffektive enn tradisjonelle prosessorer, som opererer kontinuerlig med en sentral klokke.



Det som skiller neuromorfe brikker fra klassisk databehandling, er at minne og beregning skjer på samme sted ved nevronene. Dette eliminerer den velkjente von Neumann-flaskehalsen, der data må sendes frem og tilbake mellom minne og prosessor. Kombinert med lokal læring og asynkron prosessering, gir dette svært lav latenstid og effektivitet- spesielt for edge-AI og sanntidsapplikasjoner.

Eksempler

- Intel Loihi: Forskning på spiking-nettverk for mønstergjenkjenning og robotikk, med ekstremt lavt strømforbruk.
- IBM TrueNorth: Tidlig prototype med én million nevroner, brukt i bildeanalyse og medisinsk forskning.
- BrainChip Akida: Designet for edge-AI, med støtte for både inferens og lokal læring.



Kilde: Intel Labs

Hvorfor er dette viktig?

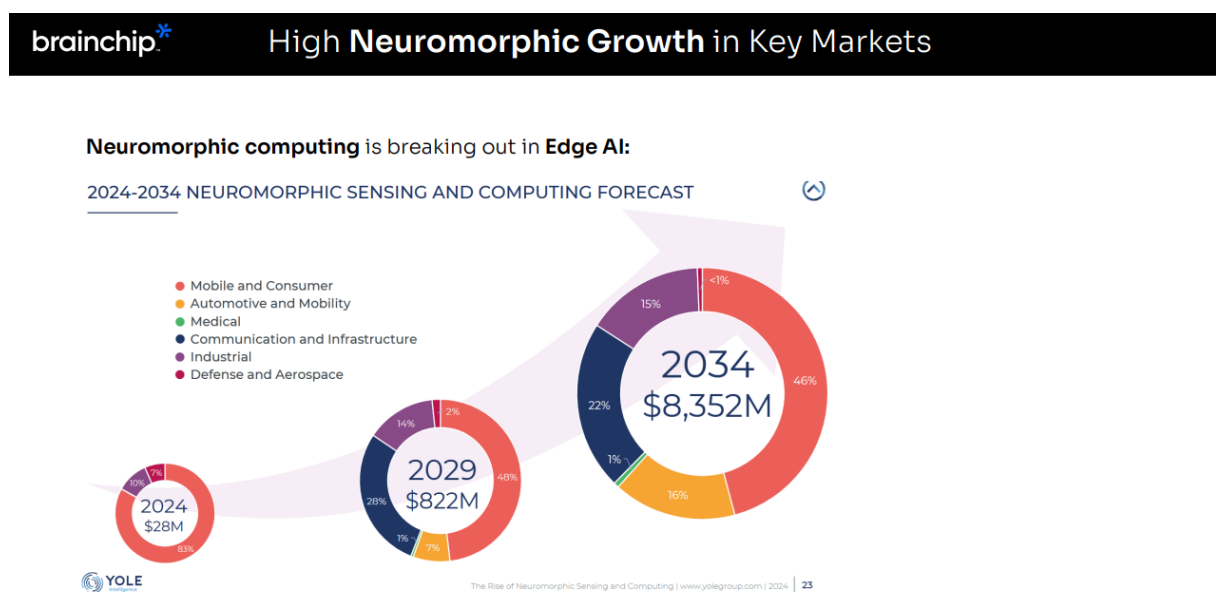
De har flere unike fordeler som gjør dem spesielt godt egnet for fremtidens teknologi. For det første er de ekstremt energieffektive. De bruker strøm bare når det faktisk skjer noe, litt som hjernen vår som bare aktiverer nevroner ved behov. Dette gjør dem perfekte for små batteridrevne enheter, som droner og sensorer, som trenger å være både lette og strømgjerrige.

I tillegg er de veldig gode på å håndtere det vi kaller «sparsomme data» i sanntid, for eksempel lyd, bilde og bevegelse, der det er viktig med rask respons og lav forsinkelse. Dette

betyr at de kan reagere umiddelbart på et lydsignal eller et visuelt mønster uten å bruke mye regnekraft.

En annen viktig fordel er at de kan lære lokalt, altså direkte på enheten, uten å sende data til skyen. Det gir både raskere tilpasning og bedre personvern, siden informasjonen aldri forlater dingsen.

Til slutt er disse brikkene skreddersydd for spesialiserte oppgaver som mønstergjenkjenning eller å kombinere informasjon fra flere sensorer, oppgaver som vanlige prosessorer ofte sliter med å gjøre effektivt.



Kilde: Brainchip IR

Selv om neuromorfe brikker har stort potensial, finnes det også noen begrensninger som gjør at de ikke passer til alle formål. For det første er de lite fleksible. De er ikke designet for generell AI-trening slik som tradisjonelle GPU-er, og passer derfor best til mer spesialiserte oppgaver.

I tillegg er økosystemet rundt teknologien fortsatt umodent. Det finnes færre utviklingsverktøy, programvarebiblioteker og støtte enn det man finner i mer etablerte AI-plattformer. Dette gjør utviklingsarbeidet mer krevende.

Det er også en høy terskel for å komme i gang. Utvikling for neuromorfe brikker krever spesialisert kunnskap, både innen nevrovitenskap og maskinvareprogrammering, noe som begrenser hvem som kan utnytte teknologien fullt ut.

Til slutt er teknologien ikke egnet for store datasett og omfattende modelltrening. Den fungerer best i såkalte edge-scenarier, altså der data behandles lokalt i små enheter, og ikke i store datasentre med tunge beregningsoppgaver.

Neuromorfe brikker hører hjemme i en mer spesialisert del av AI-infrastrukturen: nærmere kanten av nettverket enn datasenteret. De passer særlig godt i applikasjoner der sanntid, strømsparing og lokal læring er avgjørende. I det store bildet bidrar de til en mer desentralisert og bærekraftig AI-infrastruktur.

Kontekstualisering i en 2x2-matrise

Postcards From The Future

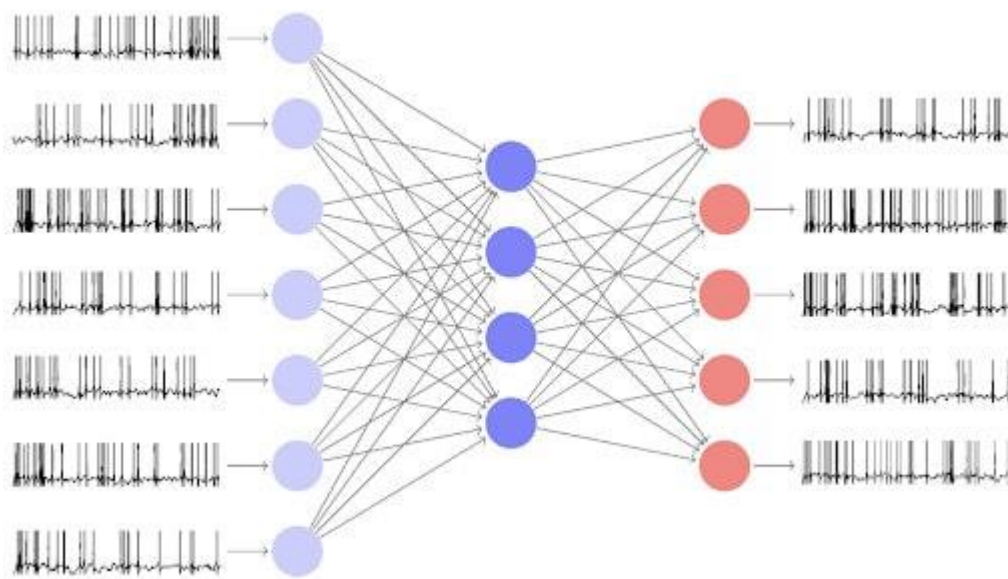
Kontekstualisering av neuromorfe brikker

		Generell	Spesialisert
Y-Akse: Anvendelighet	Sentralisert	Tradisjonelle prosessorer i datasentre, som CPUer (Intel Xeon) eller generelle servere, brukt for både AI og andre oppgaver. .	AI-optimaliserte brikker i datasentre, som GPUer og TPUer, designet for storskala modelltrening.
	Desentralisert	Generelle prosessorer for edge-enheter, som ARM-prosessorer, brukt i mobile enheter eller IoT for både AI og andre oppgaver.	Neuromorfe brikker designet for lavenergi AI-inferens på edge-enheter, som Intel Loihi eller BrainChip Akida
		X-Akse: Underliggende maskinvare	

Neuromorfe brikker representerer et paradigmeskifte i måten vi bygger AI-systemer på. De integrerer beregning og minne, reduserer energibruken, og spiller biologiske prinsipper for læring og signaloverføring. Fra et bærekraftperspektiv kan de spille en viktig rolle i å redusere karbonavtrykket innen kunstig intelligens. Samtidig er teknologien fortsatt i en tidlig fase og må modnes gjennom videre forskning, utvikling og spesialisert kompetanse.

Hva er Spiking Neural Networks?

Spiking Neural Networks (SNNs) er en type nevrale nettverk som etterligner hjernens måte å kommunisere på. I stedet for å bruke kontinuerlige signaler slik tradisjonelle nettverk gjør, opererer SNNs med diskrete elektriske pulser, såkalte «spikes», som sendes når et kunstig nevron blir tilstrekkelig aktivert. Dette gjør dem spesielt egnet for oppgaver som krever sanntidsbehandling og lavt energiforbruk, som for eksempel i autonome roboter, droner eller smarte sensorer.



Kilde: [Basic Guide to Spiking Neural Networks for Deep Learning | Intel® Tiber™ AI Studio](#)

SNNs er tett knyttet til neuromorfe brikker, som Loihi fra Intel eller Akida fra BrainChip, og er en av de AI-teknologiene vi har i dag som likner mest på hvordan den biologiske hjernen faktisk fungerer.

Dette fungerer mer som en hjerne enn en datamaskin. Hver kunstige nevron samler signaler over tid og sender først et signal videre når en terskel er nådd. Dette skjer ikke kontinuerlig, men kun når det faktisk trengs. Noe som reduserer energiforbruket dramatisk.

Læring skjer gjennom plastisitet, særlig ved hjelp av mekanismer som Spike-Timing-Dependent Plasticity (STDP). Her justeres forbindelsen mellom to nevroner basert på hvor tett i tid de sender signaler. Nettverket lærer altså gjennom mønstre i timing - ikke bare mengde.

Informasjonen i SNNs er ikke kodet i styrken på signalene, men i tidspunktet og frekvensen av «spikes». Det åpner for en helt annen form for effektiv og tidssensitiv beregning.

SNNs har flere egenskaper som gjør dem attraktive i fremtidens AI-infrastruktur:

1. *Energieffektivitet*

De er ekstremt strømgjerrige fordi nevronene kun er aktive når det skjer noe. På neuromorfe brikker kan dette bety opptil 1000 ganger lavere energiforbruk enn tradisjonelle nettverk.

2. *Sanntidsbehandling*

Siden de opererer asynkront og reagerer umiddelbart på nye signaler, passer de godt til applikasjoner som krever raske responser – som robotikk og sensorbehandling.

3. *Biologisk plausibilitet*

Fordi de etterligner hjernens signalprosessering, er SNNs ikke bare teknisk effektive – de er også et viktig verktøy i forskning på kognisjon og kunstig generell intelligens (AGI).

4. *Lokal læring*

SNNs kan lære direkte på enheten, uten å måtte sende data til skyen. Det gir lavere latenstid og bedre personvern.



Strong Market Growth for Edge AI

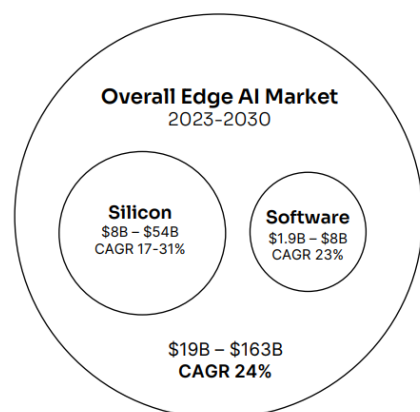
Edge AI Market is being driven by key advantages:

Why On-Device AI Matters

Bringing AI on Device Unlocks Various Benefits:



- High cloud costs are driving edge AI conversion
- Low latency requirements makes edge mandatory
- Enterprises must keep data private and secure
- Need high reliability in connection challenged environments
- Desire customization in deployment with edge learning



Dog er økosystemet i tidlig fase, med få og lite brukervennlige rammeverk. Bruksområdet er smalt og passer best til nisseoppgaver med tidsseriedata. I tillegg krever de spesialisert maskinvare, noe som gjør integrasjonen krevende for mange.

SNNs kan forstås både som en del av programvareinfrastruktur (gjennom simuleringsverktøy) og som del av beregningsinfrastruktur (når de kjører på neuromorfe brikker). De er spesielt relevante i:

- Edge-AI: Når man trenger lavt energiforbruk og rask respons – typisk i mobile enheter eller smarte sensorer.
- Sanntidsapplikasjoner: Der datastrømmen er kontinuerlig og tidssensitiv, som i autonome kjøretøy eller stemmegjenkjenning.
- Nevrovitenskapelig forskning: Der man søker å etterligne hjernen for bedre å forstå hvordan intelligens fungerer.

Vår 2x2-matrise

[Postcards From The Future](#)

Kontekstualisering av SNNs

		Generell	Spesialist
Y-Axis: Nettverket	Sentralisert	Tradisjonelle dype nevrale nettverk (DNNs) som kjører på skybaserte GPUer eller TPUer for storskala trening og inferens.	SNNs brukt i forskningsmiljøer eller simuleringsmiljøer på kraftige servere for å studere nevrovitenskap eller utvikle nye algoritmer..
	Desentr.	DNNs optimalisert for edge-enheter, som TensorFlow Lite-modeller på mobile enheter.	SNNs implementert på <u>neuromorfe</u> brikker for lavenergi-inferens på edge-enheter
		X-Axis: Anvendbarhet av teknologi	

SNNs bringer AI nærmere hjernen, både i hvordan informasjon prosesseres og hvordan læring skjer. De er et skoleeksempel på biologisk inspirert innovasjon, der nevrovitenskap møter teknologisk utvikling. Innen AI-infrastruktur danner de et spesialisert lag av systemet, optimalisert for tidsavhengige og sparsomme data, med lavt strømforbruk. Men de er fortsatt på "early adopter"-stadiet, og det kreves både teknologisk modning og utvikling av kompetanse før de får bredt fotfeste.

Noen eksempler på AI-Brikker:

IBM TrueNorth er en neuromorf brikke utviklet av IBM Research og introdusert i 2014. Den var et gjennombrudd i jakten på å bygge maskinvare som etterligner hjernen, og er basert på Spiking Neural Networks (SNNs). I stedet for å prosessere informasjon med kontinuerlige verdier slik som vanlige nevrale nettverk, bruker TrueNorth diskrete spikes, akkurat som nevroner i hjernen.

Brikken ble utviklet som del av DARPA-programmet SyNAPSE, og er designet med ett mål i sikte: ekstrem energieffektivitet. Med et strømforbruk på bare rundt 70 milliwatt for 1 million nevroner, var den et svar på utfordringen med energikrevende AI-systemer lenge før bærekraft ble et brennhett tema i teknologiverden.

Den er bygget opp som et rutenett av nevrale kjerner, 4096 i alt, der hver kjerne inneholder 256 kunstige nevroner og 256 x 256 synapser. Arkitekturen er inspirert av hjernen, med integrert minne og prosessering. Det reduserer behovet for å sende data fram og tilbake.

Brikken er asynkron og hendelses-drevet. Det betyr at prosessering kun skjer når det faktisk kommer et signal, noe som reduserer både energiforbruk og unødvendige beregninger.

Læring skjer hovedsakelig utenfor brikken, på tradisjonelle maskiner. TrueNorth er med andre ord primært en inferensmaskin – optimalisert for å bruke en ferdigtrent modell på en ekstremt strømgjerrig måte.

Hva brukes den til?

TrueNorth har først og fremst vært brukt i forskning, men har også blitt testet i praktiske applikasjoner som sanntids bilde- og lydanalyse, robotikk og sensorfusjon. Den har vist seg særlig effektiv i:

- Edge-AI, der strømforbruk og plass er begrenset
- Sensorprosessering, som å kombinere data fra kameraer, mikrofoner og andre sensorer
- Nevrovitenskap, for å simulere og forstå hjernens funksjon

Sammenlignet med Loihi og Akida

Alle tre – TrueNorth, Intel Loihi og BrainChip Akida – bruker SNNs og fokuserer på lavenergi prosessering. Men de har ulike styrker:

	TrueNorth	Loihi	Akida
Fokus	Forskning	Balanse forskning/applikasjon	Kommersialisering
Læring på chip	Begrenset	Ja (STDP og mer)	Ja
Lanseringsår	2014	2018 (Loihi 1), 2021 (Loihi 2)	2021 (Akida 1), 2023 (Akida 2)
Økosystem	NS1e, NorthPole	Lava	Integrasjon med Arm/RISC-V

TrueNorth var først ute, men Loihi og Akida har siden tatt teknologien videre og utviklet mer avanserte funksjoner og bedre integrasjon med moderne AI-verktøy.

Brikken ble introdusert i en tid da AI var dominert av GPU'er og skybasert databehandling. Den brøt med det etablerte paradigmet, og viste at det fantes andre mer energieffektive måter å bygge intelligens på.

Fra 2014 og utover representerte TrueNorth et skifte mot spesialisert maskinvare og bærekraftig AI. Mens den i hovedsak har vært en forskningsplattform, la den grunnlaget for dagens og morgendagens edge-AI.

Vår 2x2-matrise:

[Postcards From The Future](#)

TrueNorth i AI-infrastruktur

		Generell	Spesialisert
Y-Axis: Sentraliseringsgrad	Sentralt	Skybaserte systemer som bruker <u>GPUer/TPUer</u> for generelle AI-oppgaver.	<u>TrueNorth</u> brukt i forskningsmiljøer for å utvikle SNNs og utforske hjernelignende beregninger.
	Desentr. (edge)	Generelle prosessorer for edge-enheter, som ARM-brikker, brukt for både AI og andre oppgaver.	<u>TrueNorth</u> brukt på edge-enheter for lavenergi AI-inferens, sammenlignbar med <u>Loihi</u> og <u>Akida</u> .
		X-Axis: Underliggende teknologi	

TrueNorth er et tydelig eksempel på systeminnovasjon. Den utfordret rådende arkitektur med et nytt prinsipp: integrert minne og beregning. Samtidig viser den hvordan tidligfase-teknologi kan bane vei for nye paradigmer, selv om den ikke alltid blir den mest brukte.

Fra et bærekraftsperspektiv er TrueNorths energieffektivitet et varsko til en bransje med stadig høyere strømforbruk. I dag er den forbigått av nyere løsninger som Loihi 2 og Akida 2.

Intel Loihi er en neuromorf brikke utviklet av Intel Labs for å etterligne hjernens måte å prosessere informasjon på. Lansert i 2017 og videreutviklet til Loihi 2 i 2021, representerer

Loihi et av de mest ambisiøse forsøkene på å skape AI som fungerer som hjernen. Rask, tilpasningsdyktig og ekstremt energieffektiv.

The slide is titled "Introducing Loihi 2" in a large, bold, orange font. It features six key performance indicators (KPIs) in orange-bordered boxes arranged in two rows of three. The first row includes "Programmable Neurons" (described by microcode instructions), "Generalized Spikes" (spikes carry integer magnitudes for greater workload precision), and "Enhanced Learning" (support for powerful new "three factor" learning rules from neuroscience). The second row includes "10x Faster" (2-10x faster circuits² and design optimizations speed up workloads by up to 10x³), "8x More Neurons" (up to 1 million neurons per chip with up to 80x better synaptic utilization, in 1.9x smaller die), and "Better Scaling and Integration" (3D scaling with 4x more bandwidth per link⁴, >10x compression⁵ with standard interfaces). A central image shows a stack of Loihi 2 chips, with one chip highlighted in a blue frame and labeled "intel Loihi 2". Below the stack is a smaller image of a single chip. At the bottom center, a box states "Fabricated with Intel 4 process (pre-production)". The bottom left corner has the "NCL Neuromorphic Computing Lab" logo, and the bottom right corner has the "intel labs" logo. Small footnotes on the right side provide context for the performance claims.

Introducing Loihi 2

- Programmable Neurons**
Neuron models described by microcode instructions
- Generalized Spikes**
Spikes carry integer magnitudes for greater workload precision
- Enhanced Learning**
Support for powerful new "three factor" learning rules from neuroscience
- 10x Faster**
2-10x faster circuits² and design optimizations speed up workloads by up to 10x³
- 8x More Neurons**
Up to 1 million neurons per chip with up to 80x better synaptic utilization, in 1.9x smaller die
- Better Scaling and Integration**
3D scaling with 4x more bandwidth per link⁴, >10x compression⁵ with standard interfaces

Fabricated with Intel 4 process (pre-production)

² Based on silicon characterization of Loihi 1 and a combination of silicon and pre-silicon simulation estimates for Loihi 2.
³ Based on simulation modeling of a 9 layer Sigma Delta Neural Network implementation of the PilotNet DNN inference workload compared to a rate-coded SNN implementation on Loihi 1.
⁴ Based on pre-silicon circuit simulations.
⁵ Based on a 7-chip Locally Competitive Algorithm workload analysis.
See backup for analysis details. Results may vary.

NCL Neuromorphic Computing Lab intel labs

Kilde: <https://www.anandtech.com/show/16960/intel-loihi-2-intel-4nm-4>

I motsetning til tradisjonelle AI-brikker, bruker Loihi også Spiking Neural Networks (SNNs), og opererer i likhet med TrueNorth asynkront. Med Loihi 2 tok Intel teknologien et steg videre: Brikken ble raskere, mer fleksibel og enda mer energieffektiv. Og i 2024 lanserte de Hala Point, verdens største neuromorfe system, med over 1 milliard nevroner, bygget på Loihi 2, for å vise hvordan teknologien kan skaleres opp for store beregningsoppgaver.

Loihi er utviklet med tanke på fremtidens AI-systemer, og brukes allerede i forskning og utvikling av edge-AI. Aktuelle bruksområder inkluderer:

Sammenligning:

	Intel Loihi	BrainChip Akida
Fokus	Forskning og skalerbarhet	Kommersiell edge-AI
Læring på chip	Ja (STDP og adaptiv læring)	Ja (også STDP-basert)
Skalerbarhet	Loihi 2 og Hala Point	Kompakt design for IoT
Økosystem	Lava (åpen kildekode)	Arm/RISC-V-integrasjon

Selv om begge bruker SNNs, er Loihi mer forskningsrettet og skalerbar, mens Akida fokuserer på praktiske og kommersielle applikasjoner.

GPU'er og skybaserte løsninger er fortsatt dominerende, men Loihi peker mot en fremtid der AI flyttes nærmere datakilden - ut til sensorer, droner og enheter der strøm og latens betyr alt.

Fordeler:

- Svært energieffektiv
- Støtter sanntidslæring og adaptive algoritmer
- Egnet for edge-AI der skytilkobling er begrenset
- Kan skaleres til storskala systemer (Hala Point)

Begrensninger:

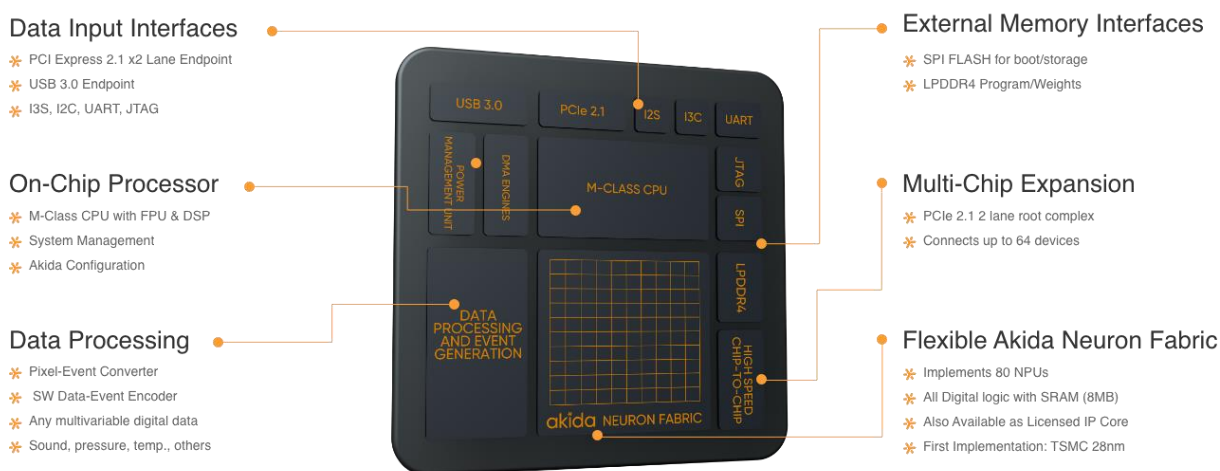
- Foreløpig mest brukt i forskning
- Økosystemet er under utvikling sammenlignet med TensorFlow
- Ikke like godt egnet for tunge språk- og bildeoppgaver
- Begrenset kommersiell utbredelse – fortsatt i tidlig fase

2x2-matrise: Loihi i AI-infrastruktur

Loihi i AI-infrastruktur

		Generell	Spesialist
Y-Axis: Anvendelighet	Sentral	Skybaserte systemer som bruker GPUer/TPUer for generelle AI-oppgaver.	Loihi 2 og Hala Point brukt i forskningsmiljøer eller storskala neuromorfe systemer for å utvikle nye algoritmer.
	Desentr. (edge)	Generelle prosessorer som ARM-brikker, brukt for både AI og andre oppgaver på edge-enheter.	Loihi brukt på edge-enheter for lavenergi, hendelsesbasert AI-inferens, sammenlignbar med BrainChip Akida.
		X-Axis: Underliggende teknologi	

Gjennom sin biologisk inspirerte tilnærming adresserer den to av AI-bransjens største utfordringer: energiforbruk og latenstid. Med støtte for lokal læring og open-source-plattformen *Lava*, inviterer Loihi forskere og utviklere til å eksperimentere og bygge nye modeller for hjernelignende AI. Samtidig befinner teknologien seg fortsatt tidlig i innovasjonsadopsjonskurven - men potensialet for å forme fremtidens AI-infrastruktur er betydelig.



BrainChip er et australsk teknologiselskap som bygger AI-prosessorer inspirert av hjernen. Deres flaggskip, **Akida**, er en neuromorf brikke designet for å gjøre kunstig intelligens ekstremt energieffektiv og kjøre direkte på enheten. Uten skytilkobling. Dette kalles edge-AI, og det er her BrainChip forsøker å skille seg ut.

Selskapet ble grunnlagt i 2004, men det var først med lanseringen av Akida i 2018 og andre generasjon i 2023 at de virkelig ble en relevant aktør innen AI-infrastruktur. Teknologien deres bygger på Spiking Neural Networks (SNNs) - der informasjon behandles som "spikes", akkurat slik biologiske nevroner gjør det. Det gir både lavt energiforbruk og mulighet for rask, lokal læring.

Hva gjør Akida unik?

I motsetning til tradisjonelle AI-brikker som trenger kontinuerlig skytilgang, lar Akida enheten lære og tilpasse seg lokalt - for eksempel ved hjelp av læringsprinsipper som STDP (Spike-Timing-Dependent Plasticity). Dette gir store fordeler i applikasjoner der personvern, sanntid og strømsparing er viktig. Som for eksempel i helsesensorer, autonome kjøretøy og industrielle IoT-systemer.

BrainChip samarbeider med store aktører som Arm og Andes Technology, og Akida kan integreres med både RISC-V og Cortex-M prosessorer. Det gjør teknologien lett å bruke for utviklere som allerede jobber med edge-enheter.

BrainChip representerer et desentralisert alternativ til dagens dominerende AI-modeller, som krever kraftige GPUer og datasentre. I stedet for å sende all data til skyen, bringer Akida intelligensen helt ut til «kanten» - nær sensoren eller brukeren. Det gjør at man kan spare energi, redusere forsinkelser og operere uten konstant nettverkstilkobling.

Viktige milepæler

- 2004: BrainChip ble grunnlagt i Perth, Australia.
- 2018: Første generasjon Akida lanseres.
- 2023: Akida 2 lanseres, med støtte for Vision Transformers og spatiotemporal AI.
- 2025: Utvidede integrasjoner med RISC-V og deltakelse på store bransjekonferanser.

Fordeler:

- Ekstremt energieffektiv, godt egnet for batteridrevne enheter
- Sanntidsprosessering og tilpasning uten skytilkobling
- Integreres lett i eksisterende økosystemer gjennom samarbeid med Arm og Andes
- Fremmer personvern, siden data ikke må sendes ut av enheten

Utfordringer:

- Neuromorf teknologi er fortsatt i en tidlig fase
- Mangler samme programvarestøtte som TensorFlow og PyTorch
- Begrenset til spesifikke bruksområder – ikke egnet for store språk- eller bildebaserte modeller
- Konkurrerer mot giganter som Intel og NVIDIA, som har større ressurser og nettverk

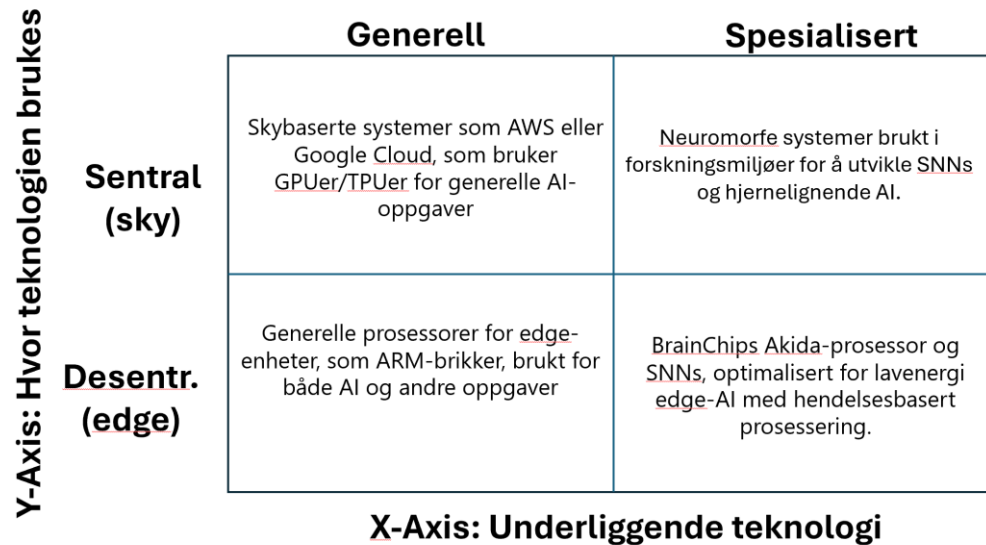
Sammenligning: BrainChip Akida vs. Intel Loihi

	BrainChip Akida	Intel Loihi
Hovedfokus	Kommersielle edge-enheter	Forskning og skalerbarhet
Læring på chip	Ja, via STDP	Ja, via STDP og andre mekanismer
Typiske applikasjoner	IoT, biler, sensorer	Roboter, nevrovitenskap, labber
Økosystem	Arm, RISC-V integrasjoner	Lava (åpen kildekode fra Intel)
Modenhet	Mer kommersiell	Mer forskningsorientert

BrainChip bygger på *disruptiv innovasjon*, der mindre og spesialiserte aktører introduserer teknologi som i første omgang virker nisjepreget, men som etter hvert kan ta over markeder når behovene endres. Dette gjelder særlig i smartbyer, autonome transportsystemer og IoT, områder der energieffektiv og edge'n er avgjørende.

2x2-matrisen vår:

BrainChip i AI-infrastruktur



BrainChip står som et godt eksempel på hvordan systemteori og innovasjonsteori møtes i praksis. Akida er ikke laget for å erstatte alle AI-halvledere, men for å fylle et spesialisert behov og gjøre det svært effektivt. Teknologien deres utfordrer de etablerte paradigmer om hvordan og hvor AI bør trenes og kjøres, og viser hvordan edge-AI og neuromorfe systemer kan bli en viktig del av en bærekraftig og personvernvennlig fremtid.

Tesla og NVIDIA

Tesla er mer enn kun en bilprodusent. Det er et teknologiselskap som utvikler autonome kjøretøy (Gjennom FSD), humanoids (Optimus) og energisystemer drevet av kunstig intelligens. I kjernen av denne satsingen ligger selskapets ambisjon om å bygge en fullstendig vertikalt integrert AI-infrastruktur. Både for trening og inferens.

NVIDIA er verdensledende innen AI-maskinvare og -programvare, kjent for sine GPU'er, SoC'er (system on chip) og verktøy som CUDA og TensorRT. NVIDIA er selve ryggraden i moderne AI-infrastruktur, både i skyen og på edge'n.

Tesla og Nvidia samarbeider tett, men også konkurrerer direkte. Denne dynamikken utfolder seg spesielt innen to sentrale deler av AI-livssyklusen: trening (modellutvikling) og handling (inferens).

Trening: Samspill og selvstendighet

Samarbeid

- **Maskinvare:** Tesla bruker NVIDIAs GPUer (H100, A100) i datasentre for å trene FSD og Optimus.
- **Programvare:** CUDA og AI Enterprise gir Tesla kraftige verktøy for optimalisert trening.

Konkurranse

- **Maskinvare:** Tesla har utviklet **Dojo**, en superdatamaskin med proprietære D1-brikker.
- **Programvare:** Tesla bygger sin egen MLOps-stack og treningsverktøy, optimalisert for Dojo.

Sett med innovasjonsteoretiske briller disrupterte NVIDIA CPU-dominansen, mens Tesla forsøker en motdisrupsjon med Dojo.

Handling (Inferens): Edge i bevegelse

Samarbeid

- **Maskinvare:** Tesla bruker NVIDIAs **DRIVE Orin** i Model S/X for sanntidsanalyse og navigasjon.
- **Programvare:** NVIDIA's TensorRT og DRIVE OS effektiviserer modellkjøring i bilene.

Konkurranse

- **Hardware:** Tesla utvikler egne SoC-er (HW3/HW4) som erstatter DRIVE Orin.
- **Software:** Teslas inferenssystemer er tilpasset deres egen maskinvare og datastrømmer.

Systemteori: NVIDIA fungerer som et støttesystem for Tesla, men Tesla bygger parallelt et uavhengig økosystem.

Tesla samler og bruker data direkte fra millioner av biler for å forbedre modellene sine. *Federated learning* gjør det mulig å oppdatere modellene uten å sende rådata til skyen.

- Samarbeid: NVIDIAs plattformer støtter federated learning og lokal prosessering.
- Konkurransen: Tesla optimaliserer HW3/HW4 for lokal læring og direkte kommunikasjon med Dojo.

Tesla får et fortrinn ved å minimere skyavhengighet og beskytte brukerdata.

Postcards From The Future

Tesla og NVIDIA i trening og handling

		Treningsmodel	Handling
Y-Axis: Samarbeidsform	Samarb.	Tesla bruker NVIDIA H100/DGX for å trene FSD- og Optimus-modeller	Tesla bruker DRIVE Orin og TensorRT for inferens i FSD
	Konkurent	Teslas Dojo konkurrerer med NVIDIA GPUer for trening.	Teslas HW3/HW4 konkurrerer med DRIVE Orin for inferens.
		X-Axis: Modell eller inferens	

Tesla konkurrerer mot	NVIDIA konkurrerer mot
NVIDIA (GPUer, DRIVE)	AMD, Google TPU, Intel Gaudi, Cerebras
Google, Amazon (Trainium)	Tesla HW3/HW4, BrainChip, Qualcomm
Qualcomm, Mobileye	Intel Loihi, Google Edge TPU

Styrker

Tesla: Full kontroll over AI-stacken, en enorm flåte med sanntidsdata og skreddersydd hardware og software

NVIDIA: Sterk utviklerlås med CUDA, er markedsleder i AI-trening, og har et økosystem med bred kundebase

Disrupsjonseksempler

Tesla	NVIDIA
Dojo utfordrer GPU-dominansen	GPU disrupterte CPU-trening (2006)
HW3/HW4 vs. DRIVE Orin	DRIVE disrupterte skybasert inferens
Federated Learning erstatter sentral dataflyt	Omdefinerte datasentre til AI-fabrikker

2x2 Matrise: Tesla vs. NVIDIA:

Postcards From The Future

Tesla og NVIDIA i trening og handling

		Treningsmodell	Handling
Y-Axis: Samarbeidsform	Samarb.	Tesla bruker NVIDIA H100/DGX for å trene FSD- og Optimus-modeller	Tesla bruker DRIVE Orin og TensorRT for inferens i FSD
	Konkurent	Teslas Dojo konkurrerer med NVIDIA GPUer for trening.	Teslas HW3/HW4 konkurrerer med DRIVE Orin for inferens.
		X-Axis: Modell eller inferens	

- **Systemteori:** NVIDIA leverer grunninfrastruktur, Tesla bygger spesifikke subsystemer.
- **Innovasjonsteori:** Samspill mellom kooperativ og disruptiv innovasjon.
- **Personvernteori:** Tesla skaper verdi ved å beholde og bearbeide data lokalt.
- **Bærekraftsteori:** Begge satser på lavere energiforbruk, men Tesla går lenger med spesialiserte brikker.

NVIDIAAs rolle i trening og handling av kunstig intelligens

Jensen Huang har tatt NVIDIA fra å være en produsent av grafikkort til å bli en hjørnestein i fremtidens økonomi. Ikke bare leverer de verdens mest etterspurte AI-brikker, men de har bygget et helt økosystem – et slags operativsystem for intelligens – som strekker seg fra datasentre til selvkjørende biler. NVIDIA er ikke lenger bare et selskap som selger chips. Det er en leverandør av infrastruktur for kunstig intelligens.

1. Trening:

AI-modeller må trenes, og trening skjer i datasentre eller "AI-fabrikker" som konverterer rådata og elektrisitet til intelligens. Her dominerer NVIDIA fullstendig.

Hardware:

Deres GPU-er (H100, A100 og nå Blackwell B200) er designet for massiv parallellprosessering, som gjør dem ideelle for å trene store språkmodeller og visuelle modeller. DGX-systemene, som kombinerer flere GPU-er med lagring og nettverk, brukes av selskaper som OpenAI, Meta og Tesla.

Software:

CUDA – NVIDIAAs programmeringsrammeverk – er en bransjestandard. Sammen med verktøy som Triton Inference Server, RAPIDS og Omniverse, gir NVIDIA utviklere og forskere verktøy for databehandling, modelltrening og simulering.

Kunder:

OpenAI trener GPT-4 med NVIDIA. Tesla bruker dem til FSD og Optimus. Google, Microsoft og Amazon tilbyr dem i skyen. NVIDIA er selve infrastrukturen.

Konkurrenter:

AMD utfordrer med MI300X, men mangler NVIDIAs økosystem. Google har TPU-er, men mindre fleksible. Startups som Cerebras og Groq er innovative, men små.

Disrupsjon:

NVIDIA erstattet CPU-er som treningsmotor med GPU-er og definerte en ny industristandard med CUDA. Nå går de videre med AI-fabrikker og syntetiske data fra Omniverse.

2. Handling:

Etter trening skal modellene ut og gjøre noe – det kalles inferens. Det kan være å tolke en pasientjournal, navigere en robot, eller kjøre en bil. Her tilbyr NVIDIA både skybaserte og edge-løsninger.

Hardware:

Jetson Orin og DRIVE Orin er spesialiserte system-on-chips (SoCs) for Edge AI – brukt i robotikk og selvkjørende kjøretøy. I skyen brukes H100 og A100 også til inferens i applikasjoner som chatboter og søk.

Software:

TensorRT optimaliserer modeller for lav latenstid og energieffektivitet. NVIDIA AI Enterprise og Triton brukes for inferens i skyen. Isaac Sim og DRIVE OS brukes til robotikk og autonom kjøring.

Kunder:

Waymo, Mercedes og Tesla bruker DRIVE for selvkjørende biler. AWS tilbyr NVIDIA GPU-er i sine skytjenester. Jetson brukes i alt fra roboter til smartkameraer.

Konkurrenter:

Qualcomm, Intel og BrainChip utfordrer med lavenergi edge-brikker. Google har Edge TPU. Men NVIDIA har fortsatt sterkest integrasjon mellom maskin og programvare.

2x2-matrise: NVIDIAs rolle i AI-infrastruktur

Postcards From The Future

NVIDIAs rolle i trening og handling

		Trening	Handling
Y-Axis: Prosessering	Sentralt	NVIDIA dominerer skybasert trening med H100, A100, og DGX-systemer.	Skybasert inferens med H100/A100 for applikasjoner som chatboter.
	Desentr. (Edge)	Federated learning støttet av DRIVE for selvkjørende biler. .	Edge AI-inferens med Jetson og DRIVE for biler og robotikk
		X-Axis: Modelltrening eller inferens	

Forklaring:

- *Trening + sentralisert*: Her dominerer NVIDIA. GPT-4 og LLaMA trenes på NVIDIA i skyen.
- *Trening + desentralisert*: Federated learning via DRIVE lar biler lære lokalt og sende tilbake innsikt.
- *Handling + sentralisert*: Apper som ChatGPT eller Amazon Alexa bruker NVIDIA i skyen.
- *Handling + desentralisert*: NVIDIA muliggjør sanntids inferens i robotikk og bilindustri med Jetson og DRIVE.

Systemteori viser NVIDIA som et sentralt subsystem i AI-økosystemet. De forbinder trening og handling, sky og edge - og gjør det via et helhetlig system av maskinvare og programvare. Innovasjonsteori forklarer hvordan NVIDIA har disruptert gamle teknologimodeller og skapt nye standarder. Med bærekraftige datasentre og støtte for *federated learning* i biler, spiller NVIDIA også en nøkkelrolle i utviklingen av ansvarlig AI.

SNNs og fremveksten av Humanoids

SNNs er ideelt egnet for Humanoids: som Tesla Optimus, Boston Dynamics' Atlas og Figure AI sine modeller.

Mens tradisjonelle AI-modeller krever enorme mengder energi og regnekraft, er SNNs utviklet for energieffektiv sanntidsprosessering direkte på enheten. Det vi kaller for Edge AI. Humanoids er ofte batteridrevne og må reagere raskt på uforutsigbare situasjoner. SNNs åpner for dette, spesielt når de kjøres på neuromorfe brikker som Intel Loihi eller BrainChip Akida.

Hvordan brukes SNNs i Humanoids?

Humanoids krever evne til å prosessere visuell, auditiv og taktil informasjon samtidig, samt til å handle adaptivt og raskt. Her gir SNNs flere fordeler:

Sensorfusjon:

SNNs håndterer input fra kameraer, LIDAR og taktile sensorer i sanntid, og gjør roboten i stand til å oppfatte og reagere på miljøet sitt med høy presisjon – samtidig som energiforbruket holdes lavt.

Motorisk kontroll:

Komplekse bevegelser som balanse, gange og gripebevegelser styres med høy respons og lav latenstid, inspirert av hvordan hjernen kontrollerer musklene.

Sosial interaksjon:

SNNs prosesserer ansiktsuttrykk, tale og kroppsspråk for å muliggjøre mer naturlig samspill med mennesker, og kan tilpasse seg ulike brukere over tid.

Lokal læring:

Gjennom mekanismer som «Spike-Timing Dependent Plasticity» (STDP) kan roboten lære og justere atferd lokalt, uten å være koblet til en sentral server.

Trening av SNNs skjer gjerne først sentralt, i datasentre. Der brukes metoder som konvertering fra tradisjonelle DNN-er, direkte spike-basert trening eller forsterkende læring i simuleringer som NVIDIA Isaac Sim. Etterpå finjusteres modellene lokalt på roboten gjennom STDP eller federated learning, hvor hver robot lærer lokalt og deler oppdateringer uten å dele persondata.

Edge-tilpasningen stiller krav til lav latenstid, robusthet, og høy energieffektivitet. Her utmerker neuromorfe brikker seg: De bruker typisk kun milliwatt, og muliggjør respons innen millisekunder. Dette er avgjørende for å unngå fall, navigere ujevnt terreng eller tolke en håndbevegelse i sanntid.

Når Humanoids lærer lokalt, må de også kunne lagre denne kunnskapen sikkert og effektivt. Dette skjer gjerne på innebygde brikker eller i kryptert lagring, og kan inkludere alt fra individuelle gangmønstre til brukerpreferanser i samtaler.

Et viktig prinsipp er at denne læringen også kan føres tilbake til sentraliserte modeller, via *federated learning* eller ved å anonymisere og aggregere data. På den måten utvikles en kollektiv intelligens basert på erfaringene til tusenvis av humanoids, uten at sensitiv informasjon forlater enhetene.

2x2-matrise: SNNs i AI-infrastruktur for Humanoids:

SNNs for humanoids i AI-infrastruktur

		Trening	Handling
Y-Axis: Prosseringen	Sentralt	Trening av SNNs i skyen eller datasentre, ofte via konvertering fra DNNs eller direkte SNN-algoritmer.	Inferens av SNNs i skyen for mindre kritiske oppgaver
	Desentr.	Federated learning for å trene SNNs på edge-humanoider, med lokale oppdateringer aggregerte sentralt	Edge AI-inferens og lokal tilpasning med SNNs på neuromorfe brikker i humanoider
		X-Axis: Modellutvikling vs inferens	

SNNs representerer en biologisk inspirert tilnærming til intelligens – skreddersydd for fremtidens roboter. Innen systemteori fungerer de som spesialiserte komponenter som knytter sammen sansing, læring og handling på en energieffektiv og autonom måte. Innovasjonsteoretisk befinner vi oss i en tidlig adopsjonsfase, men med potensial til å redefinere hvordan Edge AI utføres. Samtidig peker bærekraftsteori på fordelene ved lavt energiforbruk, og personvernteori understreker verdien av lokal læring uten datadeling.

SNNs og neuromorfe brikker gir Humanoids mulighet til å ikke bare eksistere i vår verden- men lære i den, fra den, og tilpasse seg oss.

Fra cloud til edge:

Enten vi snakker om humanoids, AI-agenter eller selvkjørende biler, handler det i bunn og grunn om én ting: læring. Å se, forstå, handle og forbedre seg over tid. Men selv om disse systemene deler felles prinsipper innen maskinlæring og dyp læring, er treningsmodellene deres skreddersydd til ulike virkeligheter. Her ser vi nærmere på hvordan modellene trenes i cloud, hvordan de flyttes ut til edge, og hvordan læring derfra strømmes tilbake – og gjør neste generasjon enda smartere.

Humanoids:

Humanoids er roboter med fysisk tilstedeværelse, laget for å etterligne menneskelig bevegelse og interaksjon. For å mestre slike ferdigheter kombineres nevrale nettverk (som SNNs og transformers) med forsterkende læring (RL) og imitasjonslæring. Bevegelse, balanse, sosial tolkning – alt læres i simuleringer som MuJoCo og Isaac Sim, hvor feil ikke koster mer enn noen millisekunder.

Flere selskaper trener humanoids på multimodal input (bilde, lyd, berøring) for å gi dem mer naturlig, adaptiv atferd. Etter treningen i cloud, overføres modellene til edge, der de må fungere med lav energi, i sanntid, og ofte uten tilkobling.

AI-agenter:

AI-agenter er digitale, ofte språkbaserte systemer som opererer i nettleseren, mobilen eller cloud. De er basert på transformer-arkitekturer og trenes som Large Language Models (LLMs), med milliarder av parametere og petabyte med tekst. Etter grunntrening finjusteres de med RLHF: Reinforcement Learning med menneskelig tilbakemelding, slik at de kan svare mer nyansert og effektivt.

De største treningsrundene skjer i cloud på tusenvis av H100-GPUer. Men i økende grad flyttes lettere versjoner av agentene til edge. For eksempel på smarttelefoner, hvor de tilpasses brukeren uten å sende data tilbake til cloud.

Selvkjørende biler:

Selvkjørende biler opererer i en fysisk, uforutsigbar verden. For å forstå omgivelser og ta beslutninger i sanntid, bruker de nevrale nettverk som SNNs og transformers, sammen med sensorfusjon og RL. Tesla, Waymo og Cruise trener modellene sine i cloud, ofte med hjelp av simuleringer og data fra virkelige kjøreturer.

Men for inferens, altså selve beslutningstakingen under kjøring, skjer alt på edge. Bilen må tolke, reagere og lære uten å vente på en cloud-respons. Derfor er modellene optimalisert for lav latency, robusthet og sikkerhet – og ofte koblet til systemer for kontinuerlig lokal læring.

Når intelligensen flyttes ut til produktene

For å flytte intelligens fra skyen til kantenheter må modellene gjøres lettere og mer effektive. Det betyr at de må forenkles og trimmes ned. Store språkmodeller trenger enklere varianter, og både SNNs og transformers må tilpasses for å kjøre raskt og strømgjerrig på spesialiserte brikker som NVIDIA Jetson, Qualcomm Snapdragon eller neuromorfe brikker som BrainChip Akida og Intel Loihi.

Edge-enheter har begrenset strøm, prosessorkraft og minne. Men de har én fordel: de er tett på virkeligheten. Og det gjør dem til de mest verdifulle læringsarenaene vi har. Gjennom federated learning sendes oppdateringer (ikke rådata) tilbake til cloud. Der kombineres de med andre enheters erfaringer for å oppdatere den globale modellen.

Tesla gjør dette med Autopilot. Google gjør det med tastaturmodeller. Apple gjør det med Siri. Det er en ny tilbakemeldingssløyfe: Cloud trener edge, edge lærer av virkeligheten og forbedrer cloud.

Edge-enheter lærer ikke bare generelt. De lærer seg deg. Din stemme, dine bevegelser og ditt miljø. Dette kalles lokalt tilpasset handling. Det skjer ofte gjennom *on-device learning*, *STDP* eller *light fine-tuning* (lett finjustering, også kalt LoRA fine-tuning).

I humanoids betyr det at roboten lærer din måte å peke på. I AI-agenter, at den forstår hvordan du formulerer deg. I biler, at de tilpasser seg veiene du bruker hver dag.

Alt skjer lokalt. Alt lagres lokalt.

En ny «feedback loop»

Trening og handling er nå ikke lenger separert. De danner en loop: Modeller trenes i cloud, distribueres til edge, lærer i sanntid, og forbedrer den sentrale modellen gjennom feedback. Resultatet er et nettverk av intelligente noder som lærer kollektivt.

Neste gang du ser en robot gå, en bil svinge av seg selv, eller en AI fullføre setningen din, husk at du er en del av læringen. Det er ikke bare maskinene som utvikler seg. Det er infrastrukturen rundt oss.

Hvor finnes de største AI-mulighetene i 2025?

For vi som følger med på fremvoksende teknologi, er spørsmålet ikke lenger *om* kunstig intelligens vil endre bransjer - men *hvor* og *hvordan*. Basert på rapporter og analyser fra blant annet McKinsey og Gartner, peker vi på syv hovedkategorier hvor AI skaper økonomisk vekst, marginforbedringer og skalerbarhet. Dette kapitlet oppsummerer disse mulighetene, kobler dem til teoretiske rammeverk, og identifiserer børsnoterte selskaper som konkret eksponerer investorer for utviklingen.

Dette er kun ment som informasjon og skal ikke anses som investeringsråd. Gjør alltid din egen analyse.

Noen muligheter i 2025

1. **AI-drevet automatisering av forretningsprosesser**

Kunstig intelligens effektiviserer repeterende oppgaver som kundeservice, regnskap og logistikk. Typiske løsninger inkluderer RPA og AI-chatbots. Dette gir høy skalering og lave variable kostnader, med svært høye bruttomarginer.

2. **Generativ AI i kreative tjenester**

AI som genererer tekst, bilder, video og musikk muliggjør verdiskaping i markedsføring, underholdning og spillutvikling. Verktøy som ChatGPT og DALL-E reduserer produksjonskost og åpner for nye forretningsmodeller.

3. **AI i cybersikkerhet**

Etter hvert som cybertrusler øker i kompleksitet, spiller AI en avgjørende rolle i trusseldeteksjon og respons i sanntid. Her har mange aktører valgt SaaS-modeller med høy margin og abonnement.

4. **AI i helsetjenester og bioteknologi**

AI gjør det mulig å analysere medisinske bilder, genomdata og symptom mønstre med

stor presisjon. Dette gir lavere kostnader, bedre diagnostikk og raskere utvikling av medisiner.

5. **AI i halvledere og infrastruktur**

Under panseret til de store språkmodellene og AI-agentene ligger spesialiserte AI-chips og datasentre. Dette er den mest kapitalkrevende, men samtidig mest uunnværlige delen av AI-økosystemet.

6. **AI for bærekraft og energioptimalisering**

Fra smarte strømmett til grønnere datasentre. AI muliggjør energieffektivisering og utslippsreduksjon. Dette skjer i takt med regulatoriske krav og økende behov for stabil, bærekraftig energibruk.

7. **AI i utdanning og læringsteknologi (EdTech)**

AI-drevet læring tilpasset hver enkelt bruker er i sterk vekst. Produkter som Duolingo viser hvordan skalerbare, digitale løsninger gir høyt brukerengasjement og lønnsomhet i læring.

Disse punktene er ikke tilfeldige. Fra et klassisk økonomisk perspektiv bygger de på skalafordeler og immaterielle eiendeler, som muliggjør høy avkastning på kapital. Innen finansanalytisk metodikk er det særlig *bruttomargin*, *inntektsvekst* og *Rule of 40* som fremhever hvilke aktører som kombinerer vekst og lønnsomhet. Og fra innovasjonsteori ser vi at mange av disse kategoriene befinner seg i den eksponentielle fasen av S-kurven, med hurtig adopsjon og betydelig verdiskaping.

Vi leter etter selskaper med høy margin, god vekst og mulig feilprising. Da finner vi en rekke børsnoterte selskaper som gir direkte eller indirekte eksponering mot AI-trendene:

- **UiPath** og **ServiceNow** (automatisering)
- **Unity Software** og **Adobe** (generativ AI)
- **Zscaler** og **SentinelOne** (cybersikkerhet)
- **Tempus** og **Intuitive Surgical** (helseteknologi)
- **NVIDIA** og **AMD** (infrastruktur og AI-chips)
- **NextEra Energy** og **Schneider Electric** (bærekraft/energi)
- **Duolingo** og **Coursera** (AI i utdanning)

Flere av disse selskapene har høy beta og er preget av negativt sentiment i markedet, men kombinerer det med sterk vekst og lønnsomhet. Det gjør dem spesielt interessante for en kontrær tilnærming, hvor markedet undervurderer fremtidig inntjening.

For å forstå hvor mulighetene er størst, har vi plassert selskapene i en 2x2-matrise basert på to dimensjoner:

1. S-kurve: Er selskapet i en eksponentiell vekstfase eller i en moden fase?
2. Feilprisings-potensial: Er markedet negativt innstilt, til tross for sterke fundamentale tall?

Postcards From The Future

Selskaper innen AI



Selskaper øverst til høyre kombinerer høy vekst og høye marginer med lav prising og negativt sentiment. Typisk der markedet overreagerer på kortsiktige problemer. Dette gjelder for eksempel Zscaler og Tempus. Nederst til venstre finner vi solide selskaper med gode marginer og lav beta, men også lav forventet oppside, som Intuitive Surgical eller Adobe.

De syv kategoriene har alle til felles at de kombinerer høy skalerbarhet med mulighet for lønnsomhet i en stadig mer automatisert økonomi. Det gir et verdifullt rammeverk for å identifisere selskaper med høy margin, vekst, og mulighet for feilprising. I en deflatorisk,

robotisert økonomi der produktivitet og software vinner terreng, kan nettopp disse aktørene vise seg å bli blant fremtidens viktigste vekstmotorer.