E-ISSN : 3025-1311
https://ejournal.warunayama.org/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 3 No 7 Tahun 2024, Hal 100-111

REVOLUSI TEKNOLOGI : TANTANGAN MASA DEPAN INTEGRASI TEKNOLOGI KECERDASAN BUATAN (AI) DALAM ARSITEKTUR KOMPUTER

Mursalin[1] , Firdaus[2], Azwa Fazilatunnisa [3], Ratna Dwi Puspita[4], Muhammad Riza Rahmatullah[5], Ann Anshori[6]

Program Studi Informatika, Fakultas Sains, UIN Sultan Maulana Hasanuddin Banten, Indonesia
Email:[1]*mursaalin90@gmail.com*, [2]*firdausonly1401@gmail.com*,
[3]*azwanisaa01@gmail.com*, [4]*puspitaratnadwi@gmail.com*,
[5]*rizarahmatullah11@gmail.com*, [6]aan.ansori@uinbanten.ac.id

Abstract
The past decade has witnessed significant advances in machine learning (ML) and artificial intelligence (AI), which have transformed fields ranging from computer vision to speech recognition and natural language processing. These advances, fueled by models such as Convolutional Neural Networks (CNN) and Reinforcement Learning (RL), have expanded AI applications in everyday life, including automotive and robotics. However, the integration of AI technologies in computer architectures faces significant challenges, mainly due to the extremely fast development speed of the ML field and the slowdown of CPU performance improvement in the post-Moore's Law era. This research aims to explore these challenges and identify potential solutions, focusing on the development of specialized hardware such as Google's Tensor Processing Unit (TPU) to accelerate learning and inference processes. In addition, this research will also explore how machine learning can aid in the chip design process itself. As such, this research seeks to provide insight into the future of AI integration in computer architecture, address existing challenges, and pave the way for the next technological revolution.

Keywords: Machine Learning, Artificial Intelligence, Computer Architecture, Tensor Processing Units, Chip Design.

Abstrak
Pada dekade terakhir kita telah menyaksikan pesatnya kemajuan dalam Machine Learning (ML) dan kecerdasan buatan (AI), yang telah mengubah berbagai bidang mulai dari visi komputer hingga pengenalan suara dan pengolahan natural language. Kemajuan ini, yang dipicu oleh model-model seperti Convolutional Neural Networks (CNN) dan Reinforcement Learning (RL), telah memperluas aplikasi AI dalam kehidupan sehari-hari, termasuk otomotif dan robotika. Namun, integrasi teknologi AI dalam arsitektur komputer menghadapi tantangan signifikan, terutama karena kecepatan perkembangan bidang ML yang sangat cepat dan perlambatan peningkatan kinerja CPU di era pasca-Hukum Moore. Penelitian ini bertujuan untuk mengeksplorasi tantangan-tantangan tersebut dan mengidentifikasi solusi potensial, dengan fokus pada pengembangan perangkat keras khusus seperti Unit Pemrosesan Tensor (TPU) Google untuk mempercepat proses pembelajaran dan inferensi. Selain itu, penelitian ini juga akan mengeksplorasi bagaimana Machine Learning dapat membantu dalam proses desain chip itu sendiri. Dengan demikian, penelitian ini berupaya untuk memberikan wawasan tentang masa depan integrasi AI dalam arsitektur komputer, mengatasi tantangan yang ada, dan membuka jalan bagi revolusi teknologi selanjutnya.

Kata Kunci: Machine Learning, Kecerdasan Buatan, Arsitektur Komputer, Unit Pemrosesan Tensor, Desain Chip.

PENDAHULUAN

Dunia telah menyaksikan kemajuan signifikan dalam bidang Machine Learning (ML) dan kecerdasan buatan (AI), khususnya, teknik deep learning yang berbasis pada jaringan saraf tiruan. Kemajuan ini telah membuka jalan baru dalam peningkatan akurasi sistem di berbagai bidang aplikasi. Menurut (Lu dkk., 2019), inovasi dalam teknologi ini telah memainkan peran kunci dalam kemajuan visi komputer, pengenalan suara, penerjemahan bahasa, dan pemahaman natural language. Penelitian-penelitian penting seperti oleh (Guo dkk., 2022; Sarker, 2021) telah menetapkan dasar bagi pengembangan lebih lanjut dalam visi komputer, sementara penelitian oleh (Nassif dkk., 2019) telah memajukan bidang pengenalan suara. Di sisi lain, pada bidang penerjemahan bahasa (Rivera Trigueros, 2022), dan berbagai tugas natural language telah ditingkatkan melalui penelitian (Ji dkk., 2021). Selain itu, Machine Learning juga telah menunjukkan potensinya dalam memecahkan tantangan yang kompleks melalui interaksi dengan lingkungan, seringkali menggunakan teknik pembelajaran penguatan.

Hal ini terbukti dari keberhasilan yang dicapai dalam berbagai bidang, mulai dari permainan strategi dalam video game seperti yang ditunjukkan oleh (X. Wang dkk., 2021). Lebih lanjut, kemajuan dalam robotika, seperti yang paparkan oleh (Cabi dkk., 2019), serta aplikasi dalam kendaraan otonom seperti yang dijelaskan oleh(Le Mero dkk., 2022), menunjukkan potensi luas dari Machine Learning dalam menavigasi dan memecahkan masalah dalam lingkungan yang kompleks. Kemajuan ini tidak hanya menandai era baru dalam teknologi AI tetapi juga membuka peluang untuk aplikasi yang lebih luas dan inovatif di masa depan. Dengan terus berkembangnya teknik dan algoritma Machine Learning, kita dapat mengantisipasi terobosan lebih lanjut yang akan terus mengubah cara kita berinteraksi dengan teknologi dan memecahkan masalah kompleks di berbagai bidang. Machine Learning, khususnya melalui deep learning dan pembelajaran penguatan, telah membuktikan dirinya sebagai salah satu bidang paling menjanjikan dan dinamis dalam penelitian AI saat ini.

Kemajuan dalam teknologi visi komputer telah mencapai titik puncaknya, terutama melalui kompetisi tahunan ImageNet Large Scale Visual Recognition Challenge (ILSVRC), yang diinisiasi oleh Fei-Fei Li dan timnya di Universitas Stanford. Kompetisi ini menantang para peneliti untuk mengembangkan algoritma yang mampu mengklasifikasikan dan mendeteksi objek dalam kumpulan data yang terdiri dari satu juta gambar berwarna yang dikategorikan ke dalam 1000 kelas yang berbeda. Sebelum era deep learning (deep learning) mendominasi kompetisi ini, metode yang digunakan cenderung mengandalkan fitur-fitur yang dirancang secara manual, dan tingkat kesalahan lima teratas (top-5 error rate) berada di atas 25% pada tahun 2010 dan 2011(Hossain & Sajib, 2019). Perubahan signifikan terjadi pada tahun 2012 ketika Alex Krizhevsky, Ilya Sutskever, dan Geoffrey Hinton dari Universitas Toronto memperkenalkan AlexNet, sebuah jaringan saraf tiruan dalam (deep neural network) yang berhasil meraih juara pertama dengan mengurangi tingkat kesalahan lima teratas menjadi 16%, sebuah pencapaian yang belum pernah terjadi sebelumnya. AlexNet menjadi satu-satunya tim yang menggunakan pendekatan jaringan saraf tiruan dalam kompetisi tersebut pada tahun itu. Keberhasilan ini memicu revolusi dalam visi komputer berbasis deep learning, dengan sebagian besar peserta pada tahun-tahun berikutnya mengadopsi pendekatan serupa, yang mengakibatkan penurunan substansial tingkat kesalahan menjadi 11,7% pada tahun berikutnya (Ismail Fawaz dkk., 2020).

Pada penelitian yang dilakukan oleh (Liu dkk., 2022; Recht dkk., 2019) menunjukkan bahwa tingkat kesalahan manusia dalam tugas ini berada di atas 5% jika dilatih selama sekitar 20 jam, atau 12% jika dilatih hanya beberapa jam, menunjukkan bahwa mesin telah mencapai dan bahkan melampaui kemampuan

manusia dalam tugas pengenalan gambar pada skala besar. Dari tahun 2011 hingga 2017, tingkat kesalahan pemenang ImageNet menurun drastis dari 26% menjadi 2,3%, menandakan kemajuan yang signifikan dalam kemampuan algoritma pengenalan gambar. Kemajuan ini tidak hanya menandai era baru dalam penelitian visi komputer tetapi juga membuka jalan bagi aplikasi praktis deep learning dalam berbagai bidang, mulai dari pengenalan wajah hingga pengembangan kendaraan otonom. Keberhasilan AlexNet dan model-model berikutnya dalam ILSVRC telah menjadi bukti pentingnya arsitektur jaringan saraf tiruan dalam dan teknik deep learning dalam mengatasi tantangan pengenalan pola pada skala besar. Ini juga menyoroti pentingnya kumpulan data yang besar dan teranotasi dengan baik, seperti ImageNet, dalam melatih model-model ini untuk mencapai tingkat akurasi yang tinggi. Seiring berjalannya waktu, kompetisi ImageNet dan pencapaian dalam deep learning terus mendorong batas-batas kemampuan mesin dalam memahami dan menginterpretasikan dunia visual dengan cara yang semakin mendekati kemampuan manusia (Recht dkk., 2019)

Meskipun kemajuan yang signifikan, tantangan utama yang dihadapi dalam integrasi teknologi kecerdasan buatan (AI) dalam arsitektur komputer adalah kecepatan luar biasa perkembangan bidang penelitian ML. Proyek desain chip dimasa ini ini membutuhkan waktu yang lama untuk menyelesaikan desain dan produksi, sementara bidang ML terus berkembang dengan cepat. Hal ini menimbulkan tantangan bagi arsitek komputer untuk memprediksi di mana bidang ML akan berada dalam jangka waktu 2 hingga 5 tahun ke depan. Selain itu, perlambatan peningkatan kinerja CPU di era pasca-Hukum Moore menuntut pengembangan perangkat keras khusus untuk Machine Learning yang dapat mengatasi kebutuhan komputasi yang semakin meningkat (Mirhoseini dkk., 2020, 2021; Settaluri dkk., 2020). Penelitian ini bertujuan untuk mengeksplorasi tantangan dan solusi dalam integrasi teknologi kecerdasan buatan dalam arsitektur komputer, dengan fokus pada pengembangan perangkat keras khusus untuk Machine Learning. Tujuan utama adalah untuk memahami bagaimana perangkat keras khusus, seperti Unit Pemrosesan Tensor (TPU) Google, dapat dirancang untuk mempercepat proses Machine Learning dan inferensi, serta bagaimana Machine Learning dapat membantu dalam proses desain chip itu sendiri. Penelitian ini juga bertujuan untuk mengidentifikasi arah masa depan dalam Machine Learning yang dapat mengintegrasikan model multi-tugas berskala besar dan jarang diaktifkan yang menggunakan perutean yang lebih dinamis, berbasis contoh dan tugas daripada model Machine Learning saat ini. Dengan demikian, penelitian ini berupaya untuk memberikan wawasan tentang bagaimana teknologi kecerdasan buatan dapat lebih terintegrasi dalam arsitektur komputer untuk mengatasi tantangan masa depan dan memajukan kemampuan sistem komputasi kita

2. METODOLOGI PENELITIAN

Penelitian ini mengadopsi pendekatan kualitatif deskriptif, dengan fokus utama pada pengumpulan data melalui studi literatur. Metode ini melibatkan penggalian teori-teori relevan dan meninjau temuan dari penelitian-penelitian sebelumnya sebagai dasar untuk mengembangkan kerangka pemikiran dalam analisis hasil penelitian. Data sekunder yang diperoleh dari studi literatur ini berasal dari berbagai sumber yang telah dipublikasikan oleh peneliti lain, termasuk buku, jurnal ilmiah, artikel, dan laporan yang berkaitan dengan topik Machine Learning, Artificial Intelligence, Arsitektur Komputer, Unit Pemrosesan Tensor (TPU), dan Desain Chip.

Dalam proses seleksi literatur, diterapkan kriteria inklusi dan eksklusi yang spesifik untuk memastikan relevansi dan aktualitas data yang dianalisis. Kriteria inklusi yang ditetapkan adalah literatur yang secara eksplisit membahas tentang Machine Learning, Artificial Intelligence, Arsitektur Komputer, Unit Pemrosesan

Tensor, dan Desain Chip. Hal ini bertujuan untuk memastikan bahwa sumber-sumber yang dipilih secara langsung berkaitan dengan fokus penelitian. Sementara itu, kriteria eksklusi yang diterapkan adalah pengecualian terhadap artikel yang diterbitkan sebelum tahun 2019, untuk menjamin bahwa analisis dilakukan berdasarkan informasi dan penemuan terkini dalam bidang yang diteliti. Proses seleksi literatur yang ketat ini memungkinkan penelitian ini untuk mengumpulkan sumber-sumber yang paling relevan dan terkini, yang akan digunakan sebagai dasar untuk analisis mendalam tentang integrasi teknologi kecerdasan buatan dalam arsitektur komputer.

3. HASIL DAN PEMBAHASAN

3.1 Hukum Moore dan Evolusi Komputasi

Hukum Moore, yang diperkenalkan oleh salah satu pendiri Intel, Gordon E. Moore, menggambarkan fenomena pertumbuhan eksponensial dalam jumlah transistor pada mikroprosesor, yang menggandakan sekitar setiap dua tahun. Fenomena ini telah menjadi pedoman utama dalam industri semikonduktor dan telah mendorong inovasi teknologi selama beberapa dekade. Namun, seiring dengan kemajuan teknologi, kita telah mencapai titik di mana peningkatan kecepatan mikroprosesor tidak lagi mengikuti prediksi Hukum Moore secara ketat. Ini sebagian besar disebabkan oleh batasan fisik dan teknis dalam pembuatan mikroprosesor yang semakin miniatur (Reuther dkk., 2019). Sebagai akibatnya, industri telah memasuki era yang dikenal sebagai Pasca Hukum Moore, di mana peningkatan kinerja komputasi tidak lagi didorong oleh peningkatan kepadatan transistor semata, tetapi juga melalui inovasi arsitektur komputasi, seperti penggunaan komputasi paralel dan pengembangan teknologi baru seperti GPU untuk tugas-tugas tertentu (Zhao dkk., 2019).

3.2 Machine Learning dan Tuntutan Komputasi

Pada awal abad ke-21, telah terjadi peningkatan minat dan kemajuan dalam bidang Machine Learning (ML) dan kecerdasan buatan (AI), khususnya dengan pengembangan deep learning dan jaringan saraf tiruan. Kemajuan ini sebagian besar dimungkinkan oleh peningkatan kinerja komputasi yang didorong oleh Hukum Moore. Namun, seiring dengan pertumbuhan eksponensial dalam data yang tersedia dan kompleksitas model ML, tuntutan terhadap sumber daya komputasi telah meningkat secara dramatis. Ini telah mendorong pengembangan dan penggunaan teknologi komputasi paralel, seperti GPU, yang menawarkan kinerja floating point yang tinggi relatif terhadap CPU, memungkinkan pelatihan model ML yang lebih besar dan lebih kompleks pada masalah dunia nyata. Lambatnya peningkatan kinerja CPU telah mendorong industri untuk mencari alternatif lain untuk memenuhi tuntutan komputasi yang meningkat dari aplikasi ML. Ini termasuk pengembangan perangkat keras khusus, seperti Tensor Processing Units (TPU) oleh Google, yang dirancang khusus untuk tugas-tugas ML, menawarkan peningkatan efisiensi dan kecepatan dalam pelatihan dan inferensi model ML (Hennessy & Patterson, 2019). Selain itu, penelitian dalam ML telah berkembang pesat, dengan lebih dari 100 artikel penelitian yang diposting setiap hari ke Arxiv, menunjukkan pertumbuhan yang signifikan dalam bidang ini (Motlagh dkk., 2023).

3.3 Perkembangan Perangkat Keras Khusus untuk Machine Learning

Pada tahun 2011 dan 2012, Google mengambil langkah awal dalam mengembangkan sistem terdistribusi yang dikenal sebagai DistBelief. Sistem ini dirancang untuk memfasilitasi pelatihan paralel dan terdistribusi jaringan neural berskala besar, menggabungkan pelatihan model dan data paralel serta pembaruan asinkron pada parameter model oleh berbagai replika komputasi. Penggunaan DistBelief memungkinkan pelatihan jaringan saraf yang lebih besar dengan dataset yang lebih luas, yang secara signifikan meningkatkan akurasi dalam pengenalan suara dan klasifikasi gambar pada pertengahan tahun 2012 (Issa dkk., 2020; Nassif dkk., 2019). Namun, tantangan muncul ketika model-

model ini harus diimplementasikan dalam sistem yang melayani ratusan juta pengguna, mengingat tuntutan komputasi yang sangat besar. Analisis awal menunjukkan bahwa untuk menerapkan sistem jaringan saraf tiruan dengan peningkatan signifikan dalam tingkat kesalahan kata untuk sistem pengenalan suara utama Google, diperlukan dua kali lipat jumlah komputer di pusat data Google. Hal ini menimbulkan pertanyaan tentang kelayakan ekonomi dan logistik, mendorong pemikiran untuk mengembangkan perangkat keras khusus untuk jaringan saraf, baik untuk inferensi maupun pelatihan (Nassif dkk., 2019).

3.4 Perangkat Keras Khusus untuk Deep learning

Model deep learning memiliki karakteristik unik yang membedakannya dari komputasi pada umumnya, termasuk toleransi terhadap presisi komputasi yang rendah, dominasi operasi aljabar linier padat, dan tidak memerlukan fitur kompleks CPU modern seperti prediksi cabang atau eksekusi spekulatif. Ini membuka peluang untuk mengembangkan perangkat keras komputasi yang dioptimalkan untuk aljabar linier padat dan presisi rendah, namun tetap dapat diprogram untuk berbagai operasi aljabar linier. Pendekatan ini mirip dengan pengembangan prosesor sinyal digital (DSP) untuk aplikasi telekomunikasi pada tahun 1980-an, tetapi dengan aplikasi yang jauh lebih luas berkat penerapan deep learning yang luas di berbagai bidang. Berdasarkan pertimbangan ini, Google memutuskan untuk memulai pengembangan serangkaian akselerator yang disebut Unit Pemrosesan Tensor (TPU) untuk mempercepat inferensi dan pelatihan deep learning. TPU pertama, TPUv1, dirancang khusus untuk akselerasi inferensi, menunjukkan peningkatan signifikan dalam kecepatan dan efisiensi energi dibandingkan dengan GPU atau CPU kontemporer[Vanhoucke et al. 2011][Jouppi et al. 2017]. TPUv1, dengan 65.536 unit multiply-accumulate 8-bit, menawarkan throughput puncak 92 TeraOps/detik (TOPS), mencapai 15X - 30X kecepatan lebih cepat dan 30X - 80X efisiensi energi lebih tinggi dibandingkan GPU atau CPU kontemporer. Ini memungkinkan Google untuk menjalankan aplikasi jaringan saraf produksi dengan keunggulan biaya dan daya yang signifikan. Selain itu, pentingnya inferensi pada perangkat seluler berdaya rendah mendorong pengembangan Edge TPU Google, yang menawarkan 4 TOPs dengan daya 2W, memungkinkan komputasi Machine Learning di perangkat tanpa memerlukan konektivitas (Gordienko dkk., 2020; Kaufman dkk., 2021; Y. E. Wang dkk., 2019)

3.5 Perkembangan Akselerator Machine Learning dan Dampaknya pada Industri

Adopsi Machine Learning (ML) yang meluas dan perannya yang semakin penting sebagai komputasi utama di dunia telah memicu ledakan dalam pengembangan akselerator baru untuk komputasi ML. Dengan lebih dari XX perusahaan rintisan yang didukung oleh modal ventura, serta perusahaan besar yang sudah mapan, industri ini menyaksikan produksi berbagai chip dan sistem baru yang dirancang khusus untuk ML. Beberapa perusahaan, seperti Cerebras, Graphcore, dan Nervana (yang telah diakuisisi oleh Intel), berfokus pada desain yang beragam untuk pelatihan ML. Sementara itu, perusahaan lain seperti Alibaba merancang chip yang lebih berfokus pada inferensi. Desain akselerator ML bervariasi, dengan beberapa menghindari penggunaan DRAM atau HBM berkapasitas besar untuk fokus pada performa tinggi pada model yang cukup kecil sehingga seluruh rangkaian parameter dan nilai tengahnya dapat ditampung dalam SRAM. Desain lainnya menyertakan DRAM atau HBM, membuatnya cocok untuk model berskala lebih besar. Cerebras, misalnya, sedang menjajaki Wafer-scale integration (WSI), sementara Edge TPU Google dirancang untuk inferensi berdaya sangat rendah di lingkungan seperti ponsel(Kimm dkk., 2021; Sakos, 2022; *Tensor Processing Units (TPUs)*, 2024).

3.6 Tantangan dalam Desain Akselerator untuk Pelatihan

Merancang perangkat keras ML yang disesuaikan untuk pelatihan, bukan hanya inferensi, merupakan upaya yang lebih kompleks. Alasannya adalah karena

sistem chip tunggal untuk pelatihan tidak dapat menyelesaikan banyak masalah dalam periode waktu yang wajar, karena satu chip saja tidak dapat menyediakan daya komputasi yang memadai. Kebutuhan untuk pelatihan model yang lebih besar pada dataset yang lebih besar berarti bahwa sering kali diperlukan penggunaan beberapa chip dalam sistem paralel atau terdistribusi. Oleh karena itu, merancang sistem pelatihan sebenarnya melibatkan merancang sistem komputer berskala lebih besar dan menyeluruh, yang membutuhkan pemikiran tentang desain chip akselerator individual serta interkoneksi berkinerja tinggi untuk membentuk superkomputer Machine Learning yang terkoneksi dengan erat. TPU generasi kedua dan ketiga Google, TPUv2 dan TPUv3, dirancang untuk mendukung pelatihan dan inferensi. Perangkat dasar individual, masing-masing terdiri dari empat chip, dirancang untuk dihubungkan bersama ke dalam konfigurasi yang lebih besar yang disebut pod. Diagram blok dari satu chip Google TPUv2 menunjukkan dua inti, dengan kapasitas komputasi utama di setiap inti yang disediakan oleh unit perkalian matriks besar yang dapat menghasilkan hasil perkalian sepasang matriks 128x128 setiap siklusnya. Setiap chip memiliki 16 GB (TPUv2) atau 32 GB (TPUv3) memori bandwidth tinggi (HBM) yang terpasang. Bentuk penyebaran TPUv3 Pod Google terdiri dari 1024 chip akselerator, yang terdiri dari delapan rak chip dan server yang menyertainya, dengan chip yang terhubung bersama dalam jaring toroidal 32x32, yang memberikan performa sistem puncak lebih dari 100 petaflop/s (*Tensor Processing Units (TPUs)*, 2024).

3.7 Format Numerik Presisi Rendah untuk Machine Learning

Akselerator Machine Learning generasi terbaru, seperti TPUv2 dan TPUv3 Google, menggunakan format floating point yang dirancang khusus yang disebut bfloat16 . Format ini berangkat dari format 16-bit setengah presisi IEEE, tetapi menyediakan format yang lebih berguna untuk Machine Learning dengan memungkinkan sirkuit pengganda yang jauh lebih murah. bfloat16 pada awalnya dikembangkan sebagai teknik kompresi lossy untuk membantu mengurangi kebutuhan bandwidth selama komunikasi jaringan bobot dan aktivasi Machine Learning dalam sistem DistBelief Google.Ternyata, komputasi Machine Learning yang digunakan dalam model deep learning lebih mementingkan rentang dinamis daripada presisi. Selain itu, salah satu area utama biaya daya sirkuit pengganda untuk format floating point adalah array penjumlah penuh yang diperlukan untuk mengalikan bagian mantissa dari dua angka input. Format IEEE fp32 membutuhkan 576 penjumlah penuh, sementara bfloat16 hanya membutuhkan 64 penjumlah penuh. Dengan membutuhkan sirkuit yang jauh lebih sedikit, format bfloat16 memungkinkan untuk menempatkan lebih banyak pengganda dalam area chip dan anggaran daya yang sama, sehingga menghasilkan akselerator ML dengan flop/detik dan flop/Watt yang lebih tinggi. Representasi presisi yang berkurang juga mengurangi bandwidth dan energi yang dibutuhkan untuk memindahkan data ke dan dari memori atau untuk mengirimkannya melintasi interkoneksi fabric, memberikan keuntungan efisiensi lebih lanjut. Format bfloat16 telah menjadi format floating andalan di TPUv2 dan TPUv3 Google sejak tahun 2015, dan pada Desember 2018, Intel mengumumkan rencana untuk menambahkan dukungan bfloat16 ke prosesor generasi mendatang. Penggunaan format numerik presisi rendah yang dirancang khusus, seperti bfloat16, menandai kemajuan penting dalam desain akselerator Machine Learning. Dengan menyediakan rentang dinamis yang memadai sambil secara signifikan mengurangi kompleksitas sirkuit, format ini memungkinkan akselerator ML yang lebih efisien secara energi dan dengan kinerja yang lebih tinggi. Adopsi luas format bfloat16 oleh pemain utama seperti Google dan Intel menunjukkan pentingnya inovasi dalam representasi numerik untuk masa depan komputasi Machine Learning yang efisien (Karthika dkk., 2020; Panda, 2019)

3.8 Tantangan dan Ketidakpastian dalam Bidang yang Berkembang Cepat

Salah satu tantangan utama dalam membangun akselerator perangkat keras untuk Machine Learning adalah kecepatan luar biasa perkembangan bidang ini. Proyek desain chip sementara ini membutuhkan waktu 18 hingga 24 bulan untuk menyelesaikan desain, fabrikasi, dan pengembalian chip, serta pemasangannya dalam lingkungan pusat data produksi. Untuk kelayakan ekonomi, chip ini harus memiliki masa pakai setidaknya tiga tahun. Oleh karena itu, arsitek komputer yang membangun perangkat keras ML dihadapkan pada tantangan memprediksi perkembangan bidang ML dalam jangka waktu 2 hingga 5 tahun ke depan. Menurut pemaparan (Chen dkk., 2022) menyatukan arsitek komputer, pengembang perangkat lunak tingkat tinggi, dan peneliti ML harus mulai mendiskusikan topik seperti "bagaimana kemungkinan perkembangan perangkat keras dalam jangka waktu tersebut?" dan "tren penelitian menarik apa saja yang mulai muncul dan implikasinya bagi perangkat keras ML?" kekritisan semacam itu merupakan cara yang berguna untuk memastikan perancangan dan pembangunan perangkat keras yang berguna untuk mempercepat penelitian dan penggunaan produksi ML. Berikut ada beberapa penelitian menarik serta impresif dimana relevan dengan tren implikatif untuk arsitek perangkat keras:

"Resiliency at Scale: Managing Google's TPUv4 Machine Learning Supercomputer" menyajikan analisis mendalam tentang desain dan pengoperasian superkomputer TPUv4 Google, dengan fokus pada pencapaian ketersediaan sistem yang tinggi dan toleransi terhadap kesalahan dalam skala besar. TPUv4, Tensor Processing Unit generasi ketiga Google, digunakan sebagai superkomputer 4096 node dengan interkoneksi torus 3D khusus. Makalah ini menyoroti tantangan penskalaan model machine learning (ML) dan menjaga ketersediaan sumber daya komputasi yang tinggi untuk pekerjaan pelatihan ML. Makalah ini memperkenalkan pendekatan software-defined networking (SDN) untuk mengelola fabric interkoneksi antar-chip (ICI) bandwidth tinggi, memanfaatkan peralihan sirkuit optik untuk konfigurasi rute dinamis guna mengatasi kegagalan. Infrastruktur ini mendeteksi kegagalan, memicu konfigurasi ulang untuk meminimalkan gangguan, dan memulai alur kerja perbaikan. Pendekatan konfigurasi ulang dinamis ini, dikombinasikan dengan strategi perutean yang toleran terhadap kesalahan, memungkinkan superkomputer TPUv4 mencapai ketersediaan sistem 99,98%, yang secara efektif menangani pemadaman perangkat keras (Zu dkk., 2024).

"TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings" membahas kemajuan dalam arsitektur TPU v4 Google, yang menekankan pada sakelar sirkuit optik (OCS) dan dukungan perangkat keras untuk penyematan. TPU v4 dihadirkan sebagai jawaban atas tuntutan beban kerja pembelajaran mesin yang terus berkembang, menawarkan peningkatan yang signifikan dalam hal skala, ketersediaan, pemanfaatan, dan performa. Pengenalan OCS memungkinkan konfigurasi ulang dinamis dari topologi interkoneksi, meningkatkan kemampuan sistem untuk mentolerir kegagalan dan mengoptimalkan beban kerja tertentu. Selain itu, makalah ini memperkenalkan SparseCores, prosesor khusus dalam TPU v4 yang memberikan akselerasi yang efisien untuk model yang mengandalkan penyematan, mencapai peningkatan kinerja yang signifikan dengan area minimal dan overhead daya. Digunakan sejak tahun 2020, TPU v4 menunjukkan kinerja dan efisiensi energi yang unggul dibandingkan dengan pendahulunya dan arsitektur khusus domain (DSA) kontemporer lainnya, menjadikannya platform yang kuat untuk melatih model pembelajaran mesin berskala besar (Jouppi dkk., 2023).

Evolusi beban kerja machine learning (ML) dan kemajuan yang terkait dalam arsitektur Tensor Processing Unit (TPU) Google mencerminkan lanskap penelitian dan aplikasi kecerdasan buatan yang berubah dengan cepat.

Pergeseran dari penggunaan multi-layer perceptron (MLP) dan convolutional neural network (CNN) ke adopsi model transformator secara luas, termasuk model bahasa besar (LLM) dan BERT, telah mendorong kebutuhan akan platform komputasi yang lebih bertenaga dan efisien. TPU v4 dari Google, dengan teknologi optical circuit switching (OCS) dan SparseCores untuk dukungan penyematan, merupakan lompatan yang signifikan dalam menjawab tantangan ini. Kemampuan TPU v4 untuk mengkonfigurasi ulang topologi interkoneksi secara dinamis dan secara efisien menangani beban kerja yang sangat berat dengan embedding menunjukkan komitmen Google untuk mengembangkan perangkat kerasnya seiring dengan kebutuhan komunitas ML. Inovasi berkelanjutan ini memastikan bahwa para peneliti dan praktisi memiliki akses ke sumber daya komputasi yang canggih, sehingga memungkinkan pengembangan model ML yang semakin kompleks dan mumpuni.

Pengenalan superkomputer TPU v4 Google menandai tonggak penting dalam bidang machine learning (ML), menawarkan kekuatan komputasi dan fleksibilitas yang belum pernah ada sebelumnya untuk melatih model yang kompleks. Sakelar sirkuit optik (OCS) TPU v4 dan SparseCores untuk dukungan penyematan menjawab tantangan kritis dalam meningkatkan beban kerja ML, termasuk kebutuhan akan ketersediaan sistem yang tinggi, penanganan penyematan yang efisien, dan kemampuan untuk beradaptasi dengan berbagai strategi paralelisme. Kemajuan teknologi ini tidak hanya mempercepat pelatihan model bahasa besar (LLM) dan model rekomendasi pembelajaran mendalam (DLRM), tetapi juga membuka jalan baru untuk penelitian dan aplikasi kecerdasan buatan. Dengan menyediakan platform komputasi yang lebih efisien dan mudah beradaptasi, TPU v4 memungkinkan para peneliti untuk mengeksplorasi model yang lebih besar dan lebih kompleks, mendorong batas-batas dari apa yang mungkin dilakukan dalam pemrosesan natural language, visi komputer, dan domain ML lainnya. Dampak TPU v4 tidak hanya di bidang akademis, tetapi juga bermanfaat bagi industri yang mengandalkan model ML mutakhir untuk berbagai aplikasi, mulai dari rekomendasi yang dipersonalisasi hingga pemahaman bahasa tingkat lanjut.

Sistem yang berjalan pada akselerator ML berskala besar dapat melatih model yang dapat melakukan ribuan atau jutaan tugas dalam satu model. Model tersebut dapat terdiri dari banyak komponen berbeda dengan struktur yang berbeda, dengan aliran data antar contoh yang relatif dinamis berdasarkan contoh per contoh. Setiap komponen mungkin menjalankan beberapa pencarian arsitektur AutoML untuk mengadaptasi strukturnya. Tugas baru dapat memanfaatkan komponen yang telah dilatih pada tugas lain jika berguna. Melalui pembelajaran multi-tugas berskala sangat besar, komponen bersama, dan perutean yang dipelajari, model ini diharapkan dapat dengan cepat belajar menyelesaikan tugas baru dengan akurasi tinggi, dengan contoh yang relatif sedikit untuk setiap tugas baru.Membangun sistem Machine Learning yang dapat menangani jutaan tugas dan belajar menyelesaikan tugas baru secara otomatis merupakan tantangan besar di bidang kecerdasan buatan dan rekayasa sistem komputer. Ini akan membutuhkan kemajuan di banyak bidang, termasuk desain sirkuit solid-state, jaringan komputer, kompiler ML, sistem terdistribusi, dan algoritme Machine Learning, untuk mendorong kecerdasan buatan ke depan dengan membangun sistem yang dapat menggeneralisasi dan menyelesaikan tugas baru secara mandiri di berbagai bidang aplikasi ML(Hennessy & Patterson, 2019; Sarker, 2021).

4. KESIMPULAN

Kemajuan signifikan Machine Learning (ML) dan kecerdasan buatan (AI) selama dekade terakhir telah memberikan dampak yang luas dan mendalam pada berbagai bidang, mulai dari ilmu pengetahuan dan teknologi hingga berbagai aspek kehidupan sehari-hari. Dengan kemajuan ini, Machine Learning

telah menjadi kekuatan pendorong di balik inovasi dan peningkatan dalam banyak sektor, memungkinkan kita untuk menyelesaikan masalah yang sebelumnya dianggap sulit atau tidak mungkin. Salah satu tantangan utama yang dihadapi industri saat ini adalah kebutuhan akan komputasi khusus yang disesuaikan dengan kebutuhan Machine Learning, terutama di tengah perlambatan peningkatan kinerja CPU untuk keperluan umum di era pasca-Hukum Moore. Era ini menandai periode di mana peningkatan kinerja komputasi tidak lagi mengikuti laju eksponensial yang diperkirakan oleh Hukum Moore, menimbulkan kebutuhan untuk mengeksplorasi pendekatan baru dalam desain perangkat keras komputasi. Dalam menghadapi tantangan ini, industri perangkat keras komputasi berada di ambang revolusi, dengan peluang untuk menerapkan serangkaian teknik baru pada berbagai masalah di banyak domain. Tujuannya adalah untuk meningkatkan skala model dan dataset yang dapat kita latih, sehingga memperluas kemampuan dan aplikasi Machine Learning.

Dengan meningkatkan skala ini, kita dapat mengharapkan untuk melihat dampak yang lebih besar dari pekerjaan ini, yang akan menyentuh hampir semua aspek kehidupan manusia. Salah satu arah yang paling menjanjikan adalah pengembangan sistem pembelajaran multi-tugas berskala besar dan masif yang dapat menggeneralisasi tugas-tugas baru. Dengan mendorong batas-batas dari apa yang mungkin dilakukan dengan teknologi ini, kita dapat menciptakan alat yang memungkinkan kita untuk mencapai lebih banyak sebagai masyarakat. Sistem semacam itu akan memungkinkan kita untuk menangani jutaan tugas secara bersamaan, belajar untuk menyelesaikan tugas baru secara otomatis, dan beradaptasi dengan kebutuhan yang terus berubah. Ini akan membuka kemungkinan baru dalam cara kita menyelesaikan masalah, membuat keputusan, dan berinovasi.Kita hidup di masa yang menarik, di mana kemajuan dalam Machine Learning tidak hanya memajukan ilmu pengetahuan dan teknologi tetapi juga memiliki potensi untuk memajukan umat manusia secara keseluruhan. Dengan terus mendorong batas-batas kemungkinan, kita dapat mengharapkan untuk melihat terobosan lebih lanjut yang akan membentuk masa depan kita dan memungkinkan kita untuk mencapai hal-hal yang belum pernah terbayangkan sebelumnya.

REFERENCES

Cabi, S., Colmenarejo, S. G., Novikov, A., Konyushkova, K., Reed, S., Jeong, R., Zolna, K., Aytar, Y., Budden, D., & Vecerik, M. (2019). Scaling data-driven robotics with reward sketching and batch reinforcement learning. *arXiv preprint arXiv:1909.12200*.

Chen, Z., Deng, Y., Wu, Y., Gu, Q., & Li, Y. (2022). Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems*, *35*, 23049–23062.

Gordienko, Y., Kochura, Y., Taran, V., Gordienko, N., Rokovyi, A., Alienin, O., & Stirenko, S. (2020). Scaling analysis of specialized tensor processing architectures for deep learning models. *Deep learning: Concepts and architectures*, 65–99.

Guo, M.-H., Xu, T.-**ng, Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R. R., Cheng, M.-M., & Hu, S.-M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, *8*(3), 331–368. https://doi.org/10.1007/s41095-022-0271-y

Hennessy, J. L., & Patterson, D. A. (2019). A new golden age for computer architecture. *Communications of the ACM*, *62*(2), 48–60.

Hossain, M. A., & Sajib, M. S. A. (2019). Classification of image using convolutional neural network (CNN). *Global Journal of Computer Science and Technology*, *19*(2), 13–14.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., & Petitjean, F. (2020). Inceptiontime:

Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, *34*(6), 1936–1962.
Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, *59*, 101894.
Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, *37*(15), 2112–2120.
Jouppi, N., Kurian, G., Li, S., Ma, P., Nagarajan, R., Nai, L., Patil, N., Subramanian, S., Swing, A., Towles, B., Young, C., Zhou, X., Zhou, Z., & Patterson, D. A. (2023). TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings. *Proceedings of the 50th Annual International Symposium on Computer Architecture*, 1–14. https://doi.org/10.1145/3579371.3589350
Karthika, P., Ma, P., & Sharif, H. (2020). *Heterogeneous Large-Scale Distributed Systems on Machine Learning* (hlm. 47–68). https://doi.org/10.4018/978-1-7998-3591-2.ch004
Kaufman, S., Phothilimthana, P., Zhou, Y., Mendis, C., Roy, S., Sabne, A., & Burrows, M. (2021). A Learned Performance Model for Tensor Processing Units. *Proceedings of Machine Learning and Systems*, *3*, 387–400.
Kimm, H., Paik, I., & Kimm, H. (2021). Performance comparision of tpu, gpu, cpu on google colaboratory over distributed deep learning. *2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, 312–319.
Le Mero, L., Yi, D., Dianati, M., & Mouzakitis, A. (2022). A survey on imitation learning techniques for end-to-end autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *23*(9), 14128–14147.
Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). *A ConvNet for the 2020s*. 11976–11986. https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html
Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks* (arXiv:1908.02265). arXiv. https://doi.org/10.48550/arXiv.1908.02265
Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J., Songhori, E., Wang, S., Lee, Y.-J., Johnson, E., Pathak, O., Bae, S., Nazi, A., Pak, J., Tong, A., Srinivasa, K., Hang, W., Tuncer, E., Babu, A., Le, Q. V., Laudon, J., … Dean, J. (2020, April 22). *Chip Placement with Deep Reinforcement Learning*. arXiv.Org. https://arxiv.longhoe.net/abs/2004.10746v1
Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J. W., Songhori, E., Wang, S., Lee, Y.-J., Johnson, E., Pathak, O., & Nazi, A. (2021). A graph placement methodology for fast chip design. *Nature*, *594*(7862), 207–212.
Motlagh, N. Y., Khajavi, M., Sharifi, A., & Ahmadi, M. (2023). The impact of artificial intelligence on the evolution of digital education: A comparative study of openAI text generation tools including ChatGPT, Bing Chat, Bard, and Ernie. *arXiv preprint arXiv:2309.02029*.
Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE access*, *7*, 19143–19165.
Panda, D. K. D. (2019). Overview of the MVAPICH project: Latest status and future roadmap. *MVAPICH2 User Group (MUG) Meeting*, *10*.
Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet? *Proceedings of the 36th International Conference on Machine Learning*, 5389–5400. https://proceedings.mlr.press/v97/recht19a.html

Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S., & Kepner, J. (2019). Survey and benchmarking of machine learning accelerators. *2019 IEEE high performance extreme computing conference (HPEC)*, 1–9.
Rivera Trigueros, I. (2022). Machine translation systems and quality assessment: A systematic review. *Language Resources and Evaluation*, *56*(2), 593–619.
Sakos, C. (2022). *Development and Evaluation with AI Tools & Devices: Google Edge TPU for General-Purpose Computing*.
Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x
Settaluri, K., Haj-Ali, A., Huang, Q., Hakhamaneshi, K., & Nikolic, B. (2020). AutoCkt: Deep Reinforcement Learning of Analog Circuit Designs. *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 490–495. https://doi.org/10.23919/DATE48585.2020.9116200
Tensor Processing Units (TPUs). (2024). Google Cloud. https://cloud.google.com/tpu
Wang, X., Song, J., Qi, P., Peng, P., Tang, Z., Zhang, W., Li, W., Pi, X., He, J., & Gao, C. (2021). SCC: An efficient deep reinforcement learning agent mastering the game of StarCraft II. *International conference on machine learning*, 10905–10915.
Wang, Y. E., Wei, G.-Y., & Brooks, D. (2019). Benchmarking TPU, GPU, and CPU platforms for deep learning. *arXiv preprint arXiv:1907.10701*.
Zhao, L., Zhou, Y., Lu, H., & Fujita, H. (2019). Parallel computing method of deep belief networks and its application to traffic flow prediction. *Knowledge-Based Systems*, *163*, 972–987.
Zu, Y., Ghaffarkhah, A., Dang, H.-V., Towles, B., Hand, S., Huda, S., Bello, A., Kolbasov, A., Rezaei, A., Du, D., Lacy, S., Wang, H., Wisner, A., Lewis, C., & Bahini, H. (2024). *Resiliency at Scale: Managing Google's TPUv4 Machine Learning Supercomputer*. https://www.usenix.org/conference/nsdi24/presentation/zu