

BUKU AJAR ARSITEKTUR KOMPUTER — BAB 15

# **BUKU AJAR ARSITEKTUR KOMPUTER**

## **BAB 15**

### **ARSITEKTUR KOMPUTER MODERN DAN TREN PERKEMBANGAN**

Buku Ajar Arsitektur Komputer 2026 | Berbasis RPS | IF221101 | 3 SKS

## IDENTITAS BAB

- Mata Kuliah: Arsitektur Komputer
- Kode: IF221101
- Bab: 15
- Fokus: arsitektur modern, heterogeneous computing, akselerator, cloud-edge, efisiensi energi, dan teknologi masa depan.

## CAPAIAN PEMBELAJARAN BAB

- Menjelaskan karakteristik arsitektur komputer modern.
- Menghubungkan CPU, memori, I/O, interkoneksi, dan paralelisme dengan sistem modern.
- Membedakan CPU, GPU, NPU, FPGA, dan ASIC.
- Menganalisis memory wall dan kebutuhan bandwidth.
- Menjelaskan cloud, edge, dan hybrid computing.
- Mengevaluasi trade-off performa, latensi, biaya, energi, dan keamanan.
- Merancang rekomendasi arsitektur untuk studi kasus modern.

## PETA KONSEP

- Evolusi Arsitektur → Heterogeneous Computing → Multicore → GPU → NPU/AI Accelerator → FPGA dan ASIC → Memory Wall → Memory-Centric Computing → Interkoneksi → Cloud → Edge → Hybrid → Energy Efficiency → Security → Quantum dan Emerging Computing.

### 15.1 Pendahuluan

Sistem komputer modern merupakan sintesis CPU, memori, I/O, interkoneksi, paralelisme, perangkat lunak, dan akselerator. Peningkatan kinerja tidak lagi hanya bergantung pada frekuensi prosesor, tetapi pada multicore, komputasi heterogen, hierarki memori, komunikasi berkecepatan tinggi, serta optimasi lintas perangkat keras dan perangkat lunak.

### 15.2 Evolusi Arsitektur Komputer

Evolusi dapat dilihat dari sistem single-core menuju multicore, multiprosesor, heterogeneous computing, accelerated computing, cloud, dan edge. Setiap tahap berusaha meningkatkan kinerja, throughput, efisiensi energi, atau kemampuan menangani kelas masalah tertentu.

### 15.3 Heterogeneous Computing

Heterogeneous computing menggunakan beberapa jenis unit komputasi dalam satu sistem. CPU menangani kontrol dan komputasi umum; GPU menangani banyak operasi data-paralel; NPU dan akselerator AI dioptimalkan untuk operasi tertentu; FPGA dapat dikonfigurasi; ASIC dirancang khusus untuk fungsi atau beban kerja tertentu.

## 15.4 CPU, GPU, NPU, FPGA, dan ASIC

CPU menekankan fleksibilitas dan alur kontrol. GPU menekankan throughput dan paralelisme data. NPU mengoptimalkan operasi neural. FPGA menawarkan fleksibilitas konfigurasi perangkat keras. ASIC menawarkan spesialisasi dan efisiensi tinggi tetapi fleksibilitas rendah. Pemilihan teknologi harus mempertimbangkan karakteristik workload.

## 15.5 Akselerator AI

Model AI modern membutuhkan operasi matriks, tensor, memori, dan komunikasi dalam skala besar. Akselerator dapat meningkatkan throughput atau efisiensi energi untuk kelas operasi tertentu. Namun, akselerator tidak menggantikan CPU secara universal karena sistem tetap membutuhkan pengendalian, koordinasi, I/O, dan perangkat lunak.

## 15.6 Memory Wall

Memory wall menggambarkan kesenjangan antara kemampuan komputasi dan kemampuan menyediakan data. Sistem modern mengatasi masalah ini melalui hierarki memori, cache, locality, bandwidth lebih tinggi, batching, kompresi, serta pendekatan yang mengurangi perpindahan data.

## 15.7 Memory-Centric dan Processing-in-Memory

Near-memory computing dan processing-in-memory membawa sebagian komputasi lebih dekat ke data. Pendekatan ini dapat mengurangi biaya data movement pada kelas masalah tertentu, tetapi menimbulkan tantangan integrasi, reliabilitas, pemrograman, akurasi, dan skalabilitas.

## 15.8 Interkoneksi Berkecepatan Tinggi

Ketika jumlah core dan akselerator meningkat, interkoneksi menjadi faktor kinerja. Bandwidth, latensi, topologi, dan arbitrase menentukan seberapa efektif CPU, memori, GPU, dan akselerator dapat bekerja bersama.

## 15.9 Cloud Computing

Cloud menggabungkan banyak server, jaringan, penyimpanan, virtualisasi, dan manajemen sumber daya. Keunggulannya meliputi kapasitas besar dan elastisitas, sedangkan tantangannya mencakup latensi, biaya, ketergantungan jaringan, serta keamanan.

## 15.10 Edge Computing

Edge computing melakukan sebagian pemrosesan dekat sumber data atau pengguna. Pendekatan ini sesuai untuk kebutuhan respons cepat, penghematan bandwidth, operasi lokal, dan kondisi ketika data tidak selalu perlu dikirim ke pusat komputasi.

## 15.11 Hybrid dan Multi-Tier Computing

Arsitektur hybrid menggabungkan cloud, on-premises, enterprise edge, dan device edge. Penempatan workload mempertimbangkan latensi, biaya, keamanan, privasi, ketersediaan, energi, dan kapasitas.

### 15.12 Efisiensi Energi dan Thermal Management

Konsumsi daya dan panas menjadi batas desain penting. Strategi meliputi manajemen daya dinamis, penggunaan akselerator yang sesuai, pengurangan perpindahan data, pengaturan sumber daya, pendinginan, dan hardware-software co-design.

### 15.13 Hardware-Software Co-Design

Optimasi modern dilakukan lintas lapisan. Algoritma, compiler, runtime, struktur data, format numerik, memori, dan akselerator perlu dirancang secara selaras. Contohnya, quantization dan sparsity dapat dipadukan dengan dukungan perangkat keras.

### 15.14 Keamanan Arsitektur

Pemisahan hak akses, isolasi memori, virtualisasi, mekanisme proteksi, dan desain perangkat keras merupakan bagian dari keamanan sistem. Kompleksitas arsitektur modern meningkatkan kemampuan sekaligus memperluas kebutuhan pengamanan.

### 15.15 Quantum dan Emerging Computing

Quantum computing menggunakan paradigma berbeda dari komputer klasik dan berpotensi unggul pada kelas masalah tertentu. Arah lain mencakup neuromorphic computing, photonic computing, dan memory-centric computing. Teknologi ini harus dipahami sebagai pelengkap atau alternatif khusus, bukan pengganti universal komputer klasik.

### 15.16 Tren Perkembangan 2026

Pengayaan teknologi 2026 menunjukkan meningkatnya heterogeneous computing, akselerator AI, optimasi data center, edge intelligence, memory-centric design, dan hardware-software co-design. Tren ini memperkuat hubungan antara konsep dasar arsitektur komputer dengan kebutuhan AI, energi, memori, dan interkoneksi modern.

### 15.17 Studi Kasus: Arsitektur AI Hybrid

Sebuah sistem menerima data real-time dan juga melatih model berukuran besar. Pemrosesan sensitif terhadap latensi dapat ditempatkan di edge, sedangkan pelatihan besar dapat menggunakan sumber daya pusat data. Mahasiswa menganalisis peran CPU, GPU/NPU, memori, penyimpanan, dan jaringan.

### 15.18 Studi Kasus: Memory Bottleneck

Sistem memiliki komputasi cepat tetapi utilisasi rendah karena data tidak tersedia cukup cepat. Analisis meliputi locality, cache miss, bandwidth memori, ukuran data, struktur data, batching, dan kemungkinan near-memory computing.

### 15.19 Aktivitas Pembelajaran

Mahasiswa membuat timeline evolusi arsitektur; membandingkan CPU, GPU, NPU, FPGA, dan ASIC; menganalisis memory bottleneck; merancang skema cloud-edge; mengidentifikasi strategi efisiensi energi; dan menyusun rekomendasi arsitektur masa depan.

### 15.20 Rangkuman

Arsitektur modern berkembang menuju multicore, paralel, terdistribusi, dan heterogen. CPU, GPU, NPU, FPGA, dan ASIC memiliki karakteristik berbeda. Memory wall menjadikan data movement sebagai faktor penting. Cloud dan edge dapat digabungkan dalam arsitektur hybrid. Efisiensi energi, thermal management, keamanan, dan hardware-software co-design merupakan pertimbangan utama. Quantum, neuromorphic, photonic, dan memory-centric computing merupakan arah pengembangan penting.

### 15.21 Glosarium

- AI Accelerator: perangkat keras yang dioptimalkan untuk operasi AI tertentu.
- ASIC: sirkuit terpadu untuk aplikasi atau fungsi khusus.
- Cloud Computing: penyediaan sumber daya komputasi melalui infrastruktur layanan.
- Edge Computing: pemrosesan dekat sumber data atau pengguna.
- FPGA: perangkat logika yang dapat dikonfigurasi.
- GPU: unit pemrosesan berorientasi throughput dan paralelisme data.
- Heterogeneous Computing: penggunaan beberapa jenis unit komputasi.
- Memory Wall: kesenjangan kemampuan komputasi dan penyediaan data.
- NPU: unit pemrosesan yang dioptimalkan untuk operasi neural.
- Processing-in-Memory: pendekatan yang membawa komputasi ke dalam atau dekat memori.
- Thermal Management: pengelolaan panas sistem.

### 15.22 Soal Latihan

#### A. Pemahaman

1. Jelaskan heterogeneous computing.
2. Bandingkan CPU dan GPU.
3. Jelaskan peran NPU dan akselerator AI.
4. Bandingkan FPGA dan ASIC.
5. Apa yang dimaksud memory wall?
6. Jelaskan processing-in-memory.
7. Bandingkan cloud dan edge computing.
8. Mengapa efisiensi energi penting?

#### B. Analisis

9. Pilih kombinasi CPU, GPU, NPU, FPGA, atau ASIC untuk tiga aplikasi berbeda.
10. Analisis sistem dengan bandwidth memori terbatas.
11. Rancang skema cloud-edge untuk kamera cerdas.
12. Bandingkan komputasi terpusat dan edge berdasarkan latensi dan privasi.
13. Usulkan strategi efisiensi energi untuk beban AI.

### C. HOTS dan Studi Kasus

14. Rancang arsitektur smart city yang menggabungkan sensor, edge, cloud, AI, dan penyimpanan.
15. Analisis pembagian workload AI untuk menurunkan biaya cloud.
16. Rancang perangkat portabel dengan batas daya dan latensi ketat.
17. Evaluasi kelayakan processing-in-memory dengan asumsi dan keterbatasan yang jelas.
18. Buat roadmap arsitektur lima tahun yang mempertimbangkan AI, keamanan, energi, dan scalability.

### 15.23 Proyek Mini: Modern Computing Architecture Designer

Mahasiswa merancang solusi arsitektur untuk sistem AI, IoT, multimedia, smart city, atau analitik data.

- Mendeskripsikan workload.
- Menentukan peran CPU, GPU, NPU, FPGA, atau ASIC.
- Merancang memori dan kebutuhan bandwidth.
- Menentukan model cloud, edge, atau hybrid.
- Membuat diagram aliran data dan interkoneksi.
- Menganalisis latensi, throughput, energi, keamanan, dan scalability.
- Menyediakan tiga alternatif desain dan matriks trade-off.

### 15.24 Daftar Pustaka

- Patterson, D. A., & Hennessy, J. L. Computer Organization and Design: The Hardware/Software Interface.
- Hennessy, J. L., & Patterson, D. A. Computer Architecture: A Quantitative Approach.
- Stallings, W. Computer Organization and Architecture: Designing for Performance.
- Null, L., & Lobur, J. The Essentials of Computer Organization and Architecture.
- IEEE Computer Society. 2026 Technology Predictions.
- Engineering Applications of Artificial Intelligence. Edge-based artificial intelligence: Understanding the evolution of hardware and software and future trends. 2026.
- ACM Computing Surveys. Generative AI at the Edge: A Comprehensive Survey of Architectures, Hardware and Applications. 2026.
- Journal of Systems Architecture. A survey on memory-centric neural network accelerators for edge intelligence. 2026.