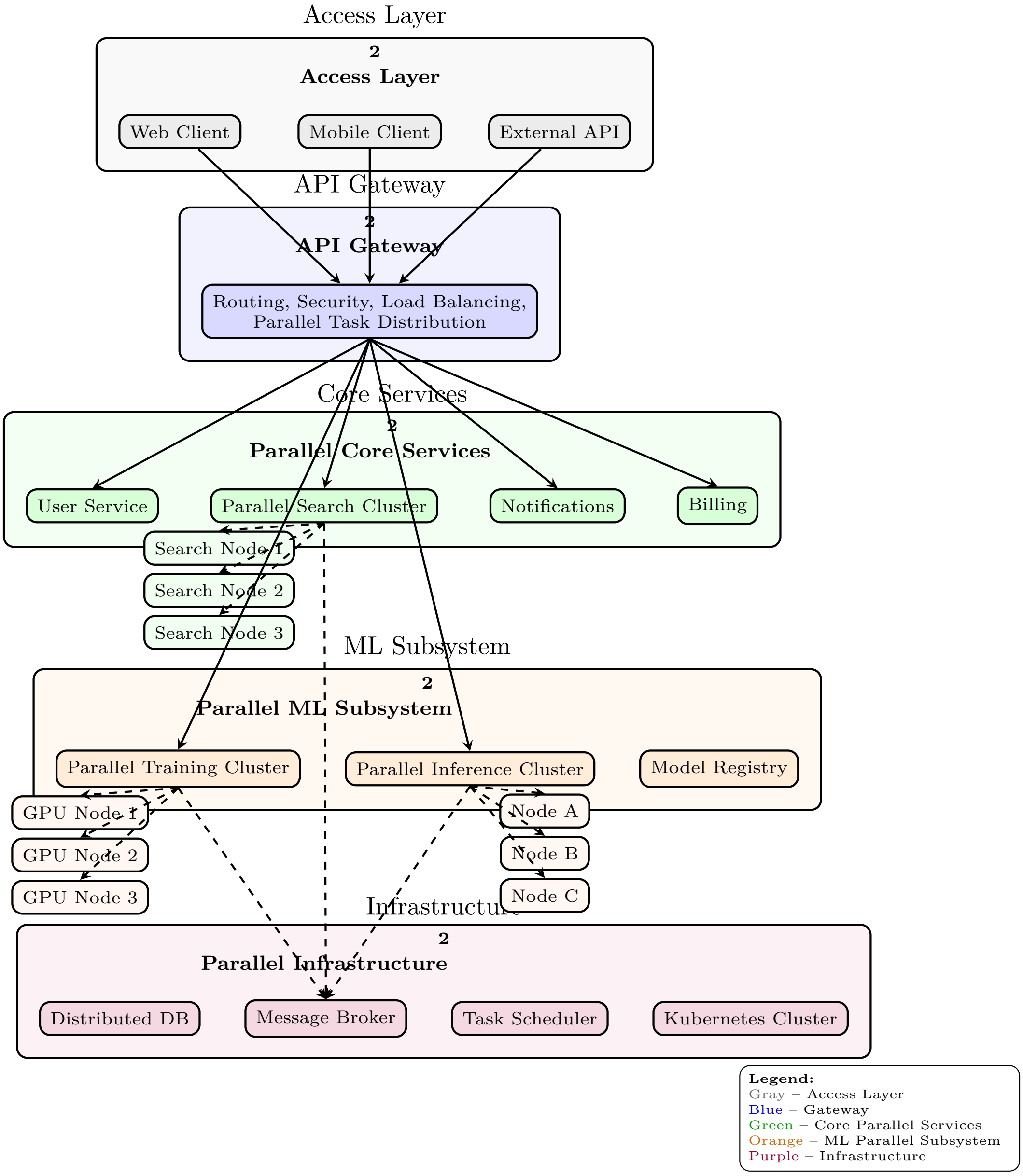
DeepSeek – Parallel Architecture

Optimized Parallel Processing Design



Description of the Parallel Architecture:

This proposed architecture transforms DeepSeek into a true *parallel computing system.* While maintaining a modular microservices organization, it introduces **intra-service parallelism** to accelerate data-intensive and AI tasks.

Access Layer: Entry point for users and APIs, ensuring uniform access. **API Gateway:** Manages authentication, load balancing, and—most importantly—*task distribution* across parallel nodes. **Parallel Core Services:** Traditional services (User, Notifications, Billing) operate normally, while the *Search Cluster* executes queries using multiple search nodes in parallel, combining their results for faster responses. **Parallel ML Subsystem:** AI tasks are split across multiple GPUs and nodes. The training and inference clusters run computations concurrently, supporting large-scale deep learning with frameworks such as Horovod or Ray. **Infrastructure:** A distributed backend (Kubernetes, broker, scheduler, and databases) coordinates job scheduling, message passing, and synchronization between parallel nodes.

By distributing workloads horizontally, DeepSeek achieves **true parallel execution**—reducing latency, improving scalability, and enabling real-time AI-driven search and analysis.