به نام خالق هوش

# AI
# Data Center
# Design Guide

# راهنمای طراحی

# سایت هوش مصنوعی

## Part 1 – V1

**Prepration by**

**Digi Hoosh HOMA**

**SABA**

**Ali Morshedsolouk**

**March 2025 | 1403 Esdand**

## فهرست مطالب

**طراحی مراکز داده‌ی هوش مصنوعی (AI Data Centers) نیازمند توجه به چندین عامل کلیدی است تا بتواند عملکرد بهینه، مقیاس‌پذیری و امنیت را تضمین کند. در ادامه به برخی از ملاحظات اصلی در طراحی این مراکز اشاره می‌شود:**

# ملاحظات اصلی طراحی یک مرکز داده هوش مصنوعی

## ۱. ظرفیت محاسباتی و پردازشی(Computing Power)

- **مراکز داده‌ی هوش مصنوعی به دلیل نیاز به پردازش حجم عظیمی از داده‌ها و اجرای مدل‌های پیچیده، به سخت‌افزارهای قدرتمند مانندGPU ها، TPUها و سرورهای پیشرفته نیاز دارند.**
- **طراحی باید به گونه‌ای باشد که امکان ارتقا و افزایش ظرفیت پردازشی در آینده وجود داشته باشد.**

## ۲. مدیریت انرژی و خنک‌سازی(Power and Cooling)

- **تجهیزات پردازشی هوش مصنوعی انرژی زیادی مصرف می‌کنند و گرمای زیادی تولید می‌کنند. بنابراین، سیستم‌های خنک‌سازی کارآمد و مدیریت انرژی بهینه از ملزومات اصلی هستند.**
- **استفاده از سیستم‌های خنک‌سازی مایع (Liquid Cooling) یا روش‌های نوین خنک‌سازی می‌تواند موثر باشد.**

## ۳. ذخیره‌سازی داده‌ها(Data Storage)

- **هوش مصنوعی به حجم بالایی از داده‌ها نیاز دارد، بنابراین سیستم‌های ذخیره‌سازی باید مقیاس‌پذیر و با دسترسی سریع (Low Latency) باشند.**
- **استفاده از فناوری‌هایی مانند ذخیره‌سازی ابری (Cloud Storage) یا سیستم‌های توزیع‌شده (Distributed Storage) می‌تواند مفید باشد.**

## ۴. شبکه‌های پرسرعت(High-Speed Networking)

- **انتقال داده‌ها بین سرورها و تجهیزات باید با کمترین تاخیر انجام شود. بنابراین، شبکه‌های پرسرعت مانند فناوری‌های ۱۰۰ GbE یا بالاتر ضروری هستند.**
- **طراحی توپولوژی شبکه باید به گونه‌ای باشد که از گلوگاه‌ها (Bottlenecks) جلوگیری کند.**

## ۵. امنیت داده‌ها(Data Security)

- **داده‌های مورد استفاده در هوش مصنوعی اغلب حساس هستند، بنابراین امنیت سایبری و محافظت از داده‌ها از اهمیت بالایی برخوردار است.**
- **استفاده از رمزنگاری(Encryption) ، فایروال‌ها و سیستم‌های تشخیص نفوذ (Intrusion Detection Systems)ضروری است.**

## ۶. مقیاس‌پذیری(Scalability)

- **مراکز داده‌ی هوش مصنوعی باید قابلیت مقیاس‌پذیری داشته باشند تا بتوانند با افزایش حجم داده‌ها و نیازهای پردازشی سازگار شوند.**
- **طراحی ماژولار و استفاده از فناوری‌های ابری می‌تواند به این امر کمک کند.**

## ۷. قابلیت اطمینان و افزونگی(Reliability and Redundancy)

- **سیستم‌های پشتیبان (Backup) و افزونگی (Redundancy) برای جلوگیری از خرابی‌ها و از دست رفتن داده‌ها ضروری هستند.**
- **استفاده از سیستم‌های UPS (منابع تغذیه بدون وقفه) و ژنراتورهای پشتیبان برای تامین انرژی مداوم مهم است.**

## ۸. بهینه‌سازی هزینه‌ها(Cost Optimization)

- **طراحی باید به گونه‌ای باشد که هزینه‌های عملیاتی و سرمایه‌ای بهینه شوند. استفاده از فناوری‌های کارآمد و کاهش مصرف انرژی می‌تواند به این امر کمک کند.**

## ۹. پایداری محیطی(Environmental Sustainability)

- **با توجه به مصرف انرژی بالا، طراحی مراکز داده‌ی هوش مصنوعی باید با توجه به پایداری محیطی انجام شود. استفاده از انرژی‌های تجدیدپذیر و کاهش ردپای کربن از جمله اهداف مهم است.**

## ۱۰. مدیریت و مانیتورینگ(Management and Monitoring)

- **سیستم‌های مدیریت و مانیتورینگ پیشرفته برای نظارت بر عملکرد تجهیزات، مصرف انرژی و امنیت ضروری هستند.**
- **استفاده از هوش مصنوعی برای مدیریت خودکار مرکز داده (AIOps) می‌تواند به بهبود کارایی کمک کند.**

**این ملاحظات به طراحی مراکز داده‌ی هوش مصنوعی کمک می‌کنند تا بتوانند به‌طور موثر از عهده‌ی نیازهای پیچیده و در حال رشد این فناوری برآیند.**

# طراحی یک مرکز داده هوش مصنوعی بر اساس نیازهای کسب و کار

**طراحی یک مرکز داده هوش مصنوعی (AI Data Center) نیازمند درک دقیق نیازها و اهداف مشتری است. برای این منظور، پرسیدن سوالات کلیدی از مشتری و استفاده از پاسخ‌ها در طراحی مرکز داده بسیار مهم است. در ادامه، برخی از سوالات اصلی و نحوه استفاده از آنها در طراحی مرکز داده هوش مصنوعی آورده شده است:**

## ۱. اهداف و نیازهای کسب‌وکار (Business Goals and Requirements)

- **سوالات:**
  - **هدف اصلی شما از راه‌اندازی مرکز داده هوش مصوعی چیست؟ (مثلاً آموزش مدل‌های بزرگ، پردازش بلادرنگ، تحلیل داده‌ها و غیره)**
  - **چه نوع کاربردهای هوش مصنوعی قرار است در این مرکز داده اجرا شود؟ (مانند پردازش زبان طبیعی، بینایی کامپیوتری، یادگیری ماشین و غیره)**
  - **آیا نیاز به پردازش بلادرنگ (Real-Time Processing) دارید؟**
- **استفاده در طراحی:**
  - **تعیین نوع سخت‌افزار مورد نیاز مانندGPU ، TPU، یا CPU .**

- طراحی سیستم‌های شبکه با تاخیر کم (Low Latency) برای کاربردهای بلادرنگ.
- انتخاب معماری مناسب برای پشتیبانی از کاربردهای خاص هوش مصوعی.

---

## ۲. حجم و نوع داده‌ها(Data Volume and Type)

- سوالات:
  - حجم داده‌هایی که قرار است پردازش شوند چقدر است؟ (مثلاً ترابایت، پتابایت یا بیشتر)
  - نوع داده‌ها چیست؟ (متن، تصویر، ویدئو، صدا و غیره)
  - آیا داده‌ها به صورت ساختاریافته، نیمه‌ساختاریافته یا غیرساختاریافته هستند؟
- استفاده در طراحی:
  - انتخاب سیستم‌های ذخیره‌سازی مناسب) مانند ذخیره‌سازی ابری، سیستم‌های توزیع‌شده یا(SAN/NAS .
  - طراحی سیستم‌های انتقال داده با پهنای باند بالا.
  - تعیین نیاز به فناوری‌های پیش‌پردازش داده(Data Preprocessing) .

---

## ۳. نیازهای پردازشی(Computing Requirements)

- سوالات:
  - چه میزان قدرت پردازشی مورد نیاز است؟) مثلاً تعدادGPU ها یاTPU ها(
  - آیا نیاز به اجرای مدل‌های یادگیری عمیق (Deep Learning) با حجم بالا دارید؟

- آیا نیاز به پردازش موازی (Parallel Processing) دارید؟

- **استفاده در طراحی:**
  - انتخاب سرورها و تجهیزات پردازشی مناسب.
  - طراحی سیستم‌های خنک‌سازی کارآمد برای مدیریت گرمای تولید شده توسط تجهیزات.
  - تعیین نیاز به منابع محاسباتی ابری (Cloud Computing) یا محلی (On-Premise).

---

## ۴. نیازهای شبکه‌ای (Networking Requirements)

- **سوالات:**
  - چه میزان پهنای باند برای انتقال داده‌ها نیاز دارید؟
  - آیا نیاز به اتصال به مراکز داده دیگر یا سیستم‌های ابری دارید؟
  - آیا نیاز به شبکه‌های با تاخیر بسیار کم (Ultra-Low Latency) دارید؟

- **استفاده در طراحی:**
  - طراحی توپولوژی شبکه‌ای مناسب) مانند(Spine-Leaf Architecture .
  - انتخاب فناوری‌های شبکه پرسرعت مانند ۱۰۰ GbEیا بالاتر.
  - اطمینان از اتصال ایمن و پایدار به منابع خارجی.

---

## ۵. نیازهای امنیتی (Security Requirements)

- **سوالات:**

- چه سطحی از امنیت برای داده‌ها و سیستم‌ها مورد نیاز است؟
- آیا داده‌های حساس یا شخصی (مانند داده‌های پزشکی یا مالی) پردازش می‌شوند؟
- آیا نیاز به رعایت استانداردهای امنیتی خاص مانندGDPR ، HIPAAو غیره دارید؟

- استفاده در طراحی:
  - پیاده‌سازی سیستم‌های رمزنگاری (Encryption) برای داده‌ها در حال انتقال و ذخیره‌شده.
  - استفاده از فایروال‌ها، سیستم‌های تشخیص نفوذ (IDS) و پیشگیری از نفوذ (IPS).
  - طراحی سیستم‌های احراز هویت و دسترسی کنترل‌شده(Access Control) .

---

## ۶. مقیاس‌پذیری و انعطاف‌پذیری(Scalability and Flexibility)

- سوالات:
  - آیا انتظار رشد حجم داده‌ها یا نیازهای پردازشی در آینده را دارید؟
  - آیا نیاز به افزودن منابع به صورت پویا (Dynamic Scaling) دارید؟

- استفاده در طراحی:
  - طراحی معماری ماژولار برای افزودن آسان منابع در آینده.
  - استفاده از فناوری‌های ابری یا ترکیبی (Hybrid Cloud) برای مقیاس‌پذیری.

---

## ۷. مدیریت انرژی و پایداری (Power Management and Sustainability)

- **سوالات:**
  - **چه میزان انرژی در دسترس است و آیا محدودیتی در مصرف انرژی وجود دارد؟**
  - **آیا نیاز به استفاده از انرژی‌های تجدیدپذیر (مانند خورشیدی یا بادی) دارید؟**
- **استفاده در طراحی:**
  - **طراحی سیستم‌های خنک‌سازی کارآمد و کاهش مصرف انرژی.**
  - **استفاده از منابع انرژی تجدیدپذیر و سیستم‌های مدیریت انرژی هوشمند.**

---

## ۸. بودجه و هزینه‌ها(Budget and Costs)

- **سوالات:**
  - **بودجه کلی شما برای طراحی و راه‌اندازی مرکز داده چقدر است؟**
  - **آیا اولویت با کاهش هزینه‌های عملیاتی (OPEX) یا سرمایه‌ای (CAPEX) است؟**
- **استفاده در طراحی:**
  - **انتخاب تجهیزات و فناوری‌هایی که با بودجه مشتری سازگار باشند.**
  - **بهینه‌سازی هزینه‌ها از طریق استفاده از فناوری‌های مقرون‌به‌صرفه یا اشتراک منابع.**

---

## ۹. قابلیت اطمینان و افزونگی(Reliability and Redundancy)

- **سوالات:**
  - **چه سطحی از دسترسی (Uptime) مورد نیاز است؟ (مثلاً ۹۹,۹٪ یا ۹۹,۹۹٪)**
  - **آیا نیاز به سیستم‌های پشتیبان (Backup) یا افزونگی (Redundancy) دارید؟**
- **استفاده در طراحی:**
  - **طراحی سیستم‌های افزونگی برای سرورها، شبکه و ذخیره‌سازی.**
  - **استفاده از سیستم‌های UPS و ژنراتورهای پشتیبان برای تامین انرژی بدون وقفه.**

---

## ۱۰. مدیریت و نظارت(Management and Monitoring)

- **سوالات:**
  - **آیا نیاز به سیستم‌های مدیریت خودکار (Automation) دارید؟**
  - **چه نوع ابزارهای مانیتورینگ و گزارش‌گیری مورد نیاز است؟**
- **استفاده در طراحی:**
  - **پیاده‌سازی سیستم‌های مانیتورینگ پیشرفته) مانندPrometheus ، Grafana).**
  - **استفاده از هوش مصنوعی برای مدیریت خودکار مرکز داده(AIOps) .**

---

**با پرسیدن این سوالات و تحلیل پاسخ‌ها، می‌توانید یک مرکز داده هوش مصنوعی طراحی کنید که به طور کامل با نیازها و اهداف مشتری شما سازگار باشد. این رویکرد تضمین می‌کند که مرکز داده نه تنها عملکرد بهینه‌ای داشته باشد، بلکه مقیاس‌پذیر، ایمن و مقرون‌به‌صرفه نیز باشد.**

# کانفیگ نمونه یک رک متوسط هوش مصنوعی

**یک رک نمونه برای کاربردهای هوش مصنوعی با پیکربندی متوسط، معمولاً شامل تجهیزات خاصی است که برای پردازش داده‌های حجیم و اجرای مدل‌های یادگیری ماشین یا یادگیری عمیق بهینه شده‌اند. در ادامه، تجهیزات معمول، مصرف انرژی و نیازهای خنک‌کننده برای چنین رکی توضیح داده شده است:**

---

**تجهیزات اصلی رک هوش مصنوعی با پیکربندی متوسط**

## ۱. سرورهای پردازشی(Compute Servers)

- **GPUسرورها :معمولاً از سرورهای مجهز به کارت‌های گرافیکی قدرتمند مانند NVIDIA A100، V100یا RTX 6000 استفاده می‌شود. این GPUها برای پردازش موازی و اجرای مدل‌های یادگیری عمیق ضروری هستند.**
- **CPUسرورها :سرورهای مجهز به پردازنده‌های قدرتمند مانند Intel Xeon یا AMD EPYCبرای مدیریت وظایف عمومی و پشتیبانی ازGPU ها.**
- **تعداد سرورها :بسته به نیاز، معمولاً ۲ تا ۴ سرور در یک رک قرار می‌گیرند.**

## ۲. ذخیره‌سازی(Storage)

- **ذخیره‌سازی سریع :(High-Performance Storage) استفاده ازSSD های NVMe با ظرفیت بالا (مثلاً ۱۰ تا ۲۰ ترابایت) برای دسترسی سریع به داده‌ها.**
- **ذخیره‌سازی شبکه‌ای :(NAS/SAN) برای ذخیره‌سازی داده‌های حجیم و دسترسی اشتراکی.**

## ۳. شبکه(Networking)

- **سوئیچ‌های پرسرعت :سوئیچ‌های ۱۰۰ GbE یا ۴۰ GbEبرای اتصال سرورها و انتقال داده‌ها با کمترین تاخیر.**
- **کارت‌های شبکه :کارت‌های شبکه پرسرعت روی سرورها برای اتصال به سوئیچ‌ها.**

## ۴. منابع تغذیه(Power Supply)

- **منابع تغذیه بدون وقفه :(UPS) برای محافظت از تجهیزات در برابر قطعی برق.**
- **PDUهای هوشمند :برای مدیریت و مانیتورینگ مصرف انرژی.**

---

## مصرف انرژی(Power Consumption)

**مصرف انرژی یک رک هوش مصنوعی با پیکربندی متوسط بسته به تعداد و نوع تجهیزات متفاوت است، اما به طور کلی:**

- **هر GPU سرور :بین ۳۰۰ تا ۱۰۰۰ وات) بسته به مدل GPU و تعداد کارت‌ها(.**
- **هر CPU سرور :بین ۲۰۰ تا ۵۰۰ وات.**
- **سوئیچ‌ها و ذخیره‌سازی :بین ۱۰۰ تا ۳۰۰ وات.**
- **کل مصرف انرژی رک :برای یک رک با ۲ تا ۴ سرور، مصرف انرژی معمولاً بین ۵ تا ۱۰ کیلووات است.**

---

## نیازهای خنک‌کننده(Cooling Requirements)

**تجهیزات هوش مصنوعی گرمای زیادی تولید می‌کنند، بنابراین سیستم‌های خنک‌کننده کارآمد ضروری هستند:**

### ۱. خنک‌کننده‌های هوا(Air Cooling)

- **فن‌های قوی :استفاده از فن‌های با جریان هوای بالا برای خنک‌کردن سرورها وGPU ها.**
- **تهویه مطبوع :(CRAC/CRAH) سیستم‌های تهویه مطبوع مخصوص دیتاسنتر برای کنترل دمای محیط.**

### ۲. خنک‌کننده‌های مایع(Liquid Cooling)

- **خنک‌کننده مستقیم مایع :(Direct Liquid Cooling) برایGPU ها وCPU ها به منظور کاهش دمای تجهیزات به طور مستقیم.**
- **خنک‌کننده مایع در رک :(In-Rack Liquid Cooling) سیستم‌های خنک‌کننده مایع که در داخل رک نصب می‌شوند.**

### ۳. محاسبه ظرفیت خنک‌کننده

- **به ازای هر کیلووات مصرف انرژی، حدود ۱,۵ تا ۲ کیلووات ظرفیت خنک‌کننده مورد نیاز است. برای مثال، یک رک با مصرف ۱۰ کیلووات به ۱۵ تا ۲۰ کیلووات ظرفیت خنک‌کننده نیاز دارد.**

---

## جمع‌بندی

- **تجهیزات اصلی GPU :سرورها، CPUسرورها، ذخیره‌سازی سریع، سوئیچ‌های پرسرعت وUPS .**
- **مصرف انرژی :بین ۵ تا ۱۰ کیلووات برای یک رک متوسط.**
- **نیازهای خنک‌کننده :سیستم‌های خنک‌کننده هوا یا مایع با ظرفیت ۱,۵ تا ۲ برابر مصرف انرژی.**

**این پیکربندی برای یک رک متوسط هوش مصنوعی مناسب است و می‌تواند بسته به نیازهای خاص مشتری (مانند حجم داده‌ها، نوع مدل‌های هوش مصنوعی و بودجه) تنظیم شود.**

# محاسبات طراحی ظرفیتهای مورد نیاز رک هوش مصنوعی

در این بخش، پارامترهای اصلی طراحی یک مرکز داده هوش مصنوعی را برمی‌شماریم و هزینه‌های اولیه و جاری هر رک را به‌صورت پارامتری بیان می‌کنیم. این رویکرد به شما کمک می‌کند تا بتوانید بر اساس نیازهای خاص خود، محاسبات را شخصی‌سازی کنید.

---

**فرمول پارامتریک طراحی مرکز داده هوش مصنوعی**

---

## 1. پارامترهای اصلی طراحی مرکز داده هوش مصنوعی

برای طراحی یک مرکز داده هوش مصنوعی، پارامترهای زیر باید در نظر گرفته شوند:

### 1.1. پارامترهای سخت‌افزاری

- **تعداد رک‌ها(N)** : تعداد رک‌های مورد نیاز برای پشتیبانی از بار کاری.
- **ظرفیت هر رکC** : تعداد سرورهایی که هر رک می‌تواند در خود جای دهد) معمولاً ۴ تا ۶ سرور برای رک‌های ۴۲ Uبا چگالی بالا(.
- **نوع سرورها** :سرورهای پایگاه داده، توسعه، هوش تجاری(BI) ، هوش مصنوعی (AI) و ذخیره‌سازی.
- **ظرفیت پردازشی هر سرور** :تعداد هسته‌هایCPU ، مقدار RAM و تعدادGPU ها.
- **ظرفیت ذخیره‌سازی هر سرور** :حجم HDD یا SSD برای ذخیره‌سازی داده‌ها و مدل‌ها.

### 1.2. پارامترهای برق و خنک‌کننده

- **برق مورد نیاز هر رکP :** برق مصرفی هر رک بر حسب کیلوولت‌آمپر(kVA) .

- **ظرفیت خنک‌کننده هر رک :(Q)** میزان خنک‌کنندگی مورد نیاز هر رک بر حسب کیلووات(kW) .
- **افزونگی** :نیاز به سیستم‌های افزونه) مانند UPS و ژنراتور (برای اطمینان از قابلیت اطمینان.

### 1.3. پارامترهای فضا

- **فضای مورد نیاز هر رک :(S)** فضای فیزیکی هر رک بر حسب متر مربع (شامل فضای راهرو برای تعمیر و نگهداری).
- **فضای کلی مرکز داده** :فضای کل مورد نیاز برای N رک.

### 1.4. پارامترهای نرم‌افزاری

- **مجوزهای نرم‌افزاری** :هزینه‌های مجوز برای سیستم‌های عامل، پایگاه‌های داده و فریم‌ورک‌های هوش مصنوعی.

---

## 2.هزینه‌های اولیه هر رک

هزینه‌های اولیه هر رک به‌صورت پارامتری به شرح زیر است:

### *2.1. هزینه‌های سخت‌افزاری هر رک*

- **سرورهای پایگاه داده** :هر سرور پایگاه داده با مشخصات ۶۴ هسته‌یCPU ، ۲۵۶ گیگابایت RAM و ۱۰ ترابایتSSD ، حدود **۲۰.۰۰۰ دلار** هزینه دارد.
- **سرورهای توسعه** :هر سرور توسعه با مشخصات ۳۲ هسته‌یCPU ، ۱۲۸ گیگابایت RAM و ۲ ترابایت SSD، حدود **۱۰.۰۰۰ دلار** هزینه دارد.
- **سرورهای هوش تجاری** :هر سرور هوش تجاری با مشخصات ۳۲ هسته‌یCPU ، ۱۲۸ گیگابایت RAMو ۲ ترابایتSSD ، حدود **۱۰.۰۰۰ دلار** هزینه دارد.
- **سرورهای هوش مصنوعی** :هر سرور هوش مصنوعی با مشخصات ۱۲۸ هسته‌یCPU ، ۵۱۲ گیگابایت RAM، ۴ x NVIDIA A100 GPUو ۲۰ ترابایتNVMe SSD ، حدود **۵۰.۰۰۰ دلار** هزینه دارد.
- **سرورهای ذخیره‌سازی** :هر سرور ذخیره‌سازی با مشخصات ۶۴ هسته‌یCPU ، ۲۵۶ گیگابایت RAM و ۱۰۰ ترابایتHDD ، حدود **۱۵.۰۰۰ دلار** هزینه دارد.

### 2.2. هزینه‌های برق و خنک‌کننده هر رک

- **سیستم UPS**: هر رک به حدود **٦ کیلوولت‌آمپر** برق نیاز دارد. هزینه‌ی UPS برای هر رک حدود **۳.۰۰۰ دلار** است.
- **ژنراتور**: هزینه‌ی ژنراتور برای هر رک حدود **٤.۰۰۰ دلار** است.
- **سیستم خنک‌کننده**: هر رک به حدود **۲۰ کیلووات** خنک‌کنندگی نیاز دارد. هزینه‌ی سیستم خنک‌کننده برای هر رک حدود **۵.۰۰۰ دلار** است.

### 2.3. هزینه‌های فضا و زیرساخت هر رک

- **فضای فیزیکی**: هر رک به حدود **۲.۵ متر مربع** فضا نیاز دارد. هزینه‌ی ساخت و آماده‌سازی فضا برای هر رک حدود **۱۰.۰۰۰ دلار** است.
- **رک و کابل‌کشی**: هزینه‌ی رک و کابل‌کشی برای هر رک حدود **۲.۰۰۰ دلار** است.

### 2.4.هزینه‌های نرم‌افزاری هر رک

- **مجوزهای نرم‌افزاری**: هزینه‌ی مجوزهای نرم‌افزاری برای هر رک حدود **۲.۰۰۰ دلار** است.

---

## 3. هزینه‌های جاری ماهیانه هر رک

هزینه‌های جاری ماهیانه هر رک به‌صورت پارامتری به شرح زیر است:

### 3.1. هزینه‌های برق

- **مصرف برق**: هر رک حدود **٦ کیلوولت‌آمپر** برق مصرف می‌کند. هزینه‌ی برق برای هر رک حدود **۵۰۰ دلار** در ماه است.

### 3.2. هزینه‌های اینترنت و شبکه

- **اتصال اینترنت**: هزینه‌ی اینترنت پرسرعت برای هر رک حدود **۱۰۰ دلار** در ماه است.

### 3.3. هزینه‌های نگهداری

- **نگهداری سخت‌افزار**: هزینه‌ی نگهداری سخت‌افزار برای هر رک حدود **۳۰۰ دلار** در ماه است.

- **نگهداری نرم‌افزار** :هزینه‌ی نگهداری نرم‌افزار برای هر رک حدود **۱۰۰ دلار** در ماه است.

### *3.4. هزینه‌های متفرقه*

- **لوازم اداری و سایر هزینه‌ها** :هزینه‌های متفرقه برای هر رک حدود **۱۰۰ دلار** در ماه است.

---

## 4. فرمول‌های محاسباتی

### *4.1. هزینه‌های اولیه کل*

هزینه‌های اولیه کل=$N$ *ضربدر*

(هزینه‌های سخت‌افزاری+هزینه‌های برق و خنک‌کننده+هزینه‌های فضا و زیرساخت+هزینه‌های نرم‌افزاری)

### *4.2. هزینه‌های جاری ماهیانه کل*

هزینه‌های جاری ماهیانه کل= $N$ x (هزینه‌های برق+هزینه‌های اینترنت+هزینه‌های نگهداری+هزینه‌های متفرقه)

(هزینه‌های برق+هزینه‌های اینترنت+هزینه‌های نگهداری+هزینه‌های متفرقه)

---

## 5. مثال محاسباتی

فرض کنید می‌خواهید یک مرکز داده با ۱۰ رک راه‌اندازی کنید:

### *5.1.هزینه‌های اولیه*

{هزینه‌های اولیه کل} = ۱۰) × ۵۰.۰۰۰ + ۳.۰۰۰ + ۴.۰۰۰ + ۵.۰۰۰ + ۱۰.۰۰۰ + ۲.۰۰۰ + ۲.۰۰۰)

= ۱۰ × ۷۶.۰۰۰ = ۷۶۰.۰۰۰ دلار

### 5.2.هزینه‌های جاری ماهیانه

{هزینه‌های جاری ماهیانه کل} = ۱۰) × ۵۰۰ + ۱۰۰ + ۳۰۰ + ۱۰۰ + ۱۰۰ = (۱۰ × ۱.۱۰۰ = ۱۱.۰۰۰ } دلار{

---

## 5. مثال محاسباتی

فرض کنید می‌خواهید یک مرکز داده با ۱۰ رک راه‌اندازی کنید:

### 5.1. هزینه‌های اولیه

هزینه‌های اولیه کل=۱۰×(۵۰.۰۰۰+۳.۰۰۰+۴.۰۰۰+۵.۰۰۰+۱۰.۰۰۰+۲.۰۰۰+۲.۰۰۰)=۱۰×۷۶.۰۰۰=۷۶۰.۰۰۰ دلار

### 5.2. هزینه‌های جاری ماهیانه

هزینه‌های جاری ماهیانه کل=۱۰)×۵۰۰+۱۰۰+۳۰۰+۱۰۰+۱۰۰=(۱۰×۱.۱۰۰=۱۱.۰۰۰ دلار

---

## 6. خلاصه بخش

با استفاده از این رویکرد پارامتریک، می‌توانید به‌راحتی هزینه‌های اولیه و جاری مرکز داده هوش مصنوعی خود را بر اساس تعداد رک‌ها و نیازهای خاص خود محاسبه کنید. این روش به شما کمک می‌کند تا تصمیم‌گیری‌های دقیق‌تری در مورد طراحی و راه‌اندازی مرکز داده داشته باشید.

# طراحی ظرفیت سنجی و GPU مورد نیاز بر اساس محاسبات نیازهای کاربر و درخواستهای اپ هوش مصنوعی

در این بخش ما بر اساس پارامترهای درخواست‌های نرم‌افزار هوش مصنوعی (AI) از طرف اپلیکیشن‌های هوش مصنوعی طراحی را فرموله می‌کنیم. این پارامترها شامل مواردی مانند **همزمانی درخواست‌ها**، **نوع درخواست‌ها** (مانند استنتاج، آموزش، پردازش زبان طبیعی، بینایی کامپیوتری و غیره)، و سایر عوامل تأثیرگذار بر تعدادGPU ها و میزان RAM مورد نیاز هستند.

---

## 1. پارامترهای اصلی درخواست‌های هوش مصنوعی

### 1.1. پارامترهای مربوط به درخواست‌ها

- **تعداد درخواست‌های همزمانR :** تعداد درخواست‌هایی که به‌طور همزمان به سیستم ارسال می‌شوند.
- **نوع درخواست :(T)** نوع درخواست هوش مصنوعی (مانند استنتاج، آموزش، پردازش زبان طبیعی، بینایی کامپیوتری و غیره).
- **حجم داده هر درخواست :(D)** حجم داده‌ای که هر درخواست نیاز دارد (بر حسب مگابایت یا گیگابایت).
- **زمان پردازش هر درخواستP :** زمان مورد نیاز برای پردازش هر درخواست (بر حسب ثانیه).

### 1.2. پارامترهای مربوط به منابع سخت‌افزاری

- **تعدادGPU های مورد نیاز :(G)** تعدادGPU هایی که برای پردازش درخواست‌ها لازم است.
- **میزان RAM مورد نیاز :(M)** مقدار حافظه‌ی RAM مورد نیاز برای پردازش درخواست‌ها (بر حسب گیگابایت).

- **ظرفیت پردازشی هر GPU یاC** : ظرفیت پردازشی هر GPU بر حسب) TFLOPS ترا فلاپس(.

---

## 2. فرمول‌های محاسباتی

### 2.1.محاسبه‌ی تعدادGPU های مورد نیاز(G)

$$G = \left\lceil \frac{R \times P \times F_T}{C \times U} \right\rceil$$

- **R**: تعداد درخواست‌های همزمان.
- **P**: زمان پردازش هر درخواست (ثانیه).
- **F_T**: ضریب پیچیدگی نوع درخواست (مثلاً ۱ برای استنتاج ساده، ۲ برای پردازش زبان طبیعی، ۳ برای بینایی کامپیوتری و غیره).
- **C**: ظرفیت پردازشی هرGPU (TFLOPS) .
- **U**: ضریب استفاده ازGPU (معمولاً بین ۰.۷ تا ۰.۹).

### 2.2. محاسبه‌ی میزان RAM مورد نیاز(M)

$$M = R \times D \times K_T$$

M = R × D × K_T

- **R**:تعداد درخواست‌های همزمان.
- **D**:حجم داده هر درخواست (گیگابایت).

- **K_T**: ضریب حافظه‌ی مورد نیاز برای نوع درخواست (مثلاً ۱.۵ برای استنتاج، ۲ برای آموزش، ۲.۵ برای پردازش زبان طبیعی و غیره).

---

# 3. حالت‌های مختلف فرمول‌ها

## 3.1. حالت‌های مختلف برای تعدادGPU ها

- **حالت ۱: استنتاج ساده(Inference)**
  - **F_T= ۱**
  - **مثال**: اگر R = ۱۰۰، P = ۰.۱ ثانیه، C = ۱۰TFLOPS ، U = ۰.۸:

$$G = \left\lceil \frac{۱۰۰ \times ۰.۱ \times ۱}{۱۰ \times ۰.۸} \right\rceil = \left\lceil \frac{۱۰}{۸} \right\rceil = ۲$$

- **حالت ۲: پردازش زبان طبیعی(NLP)**
  - **F_T= ۲**
  - **مثال**: اگر R = ۱۰۰، P= ۰.۲ ثانیه، C= ۱۰TFLOPS ، U= ۰.۸:

$$G = \left\lceil \frac{۱۰۰ \times ۰.۲ \times ۲}{۱۰ \times ۰.۸} \right\rceil = \left\lceil \frac{۴۰}{۸} \right\rceil = ۵$$

- **حالت ۳: بینایی کامپیوتری(Computer Vision)**
  - **F_T= ۳**
  - **مثال**: اگر R = ۱۰۰، P= ۰.۳ ثانیه، C= ۱۰TFLOPS ، U= ۰.۸:

$$G = \left\lceil \frac{۱۰۰ \times ۰.۳ \times ۳}{۱۰ \times ۰.۸} \right\rceil = \left\lceil \frac{۹۰}{۸} \right\rceil = ۱۲$$

## 3.2. حالت‌های مختلف برای میزانRAM

- **حالت ۱: استنتاج ساده(Inference)**
  - **۱.۵K_T =**
  - **مثال** :اگر R = ۱۰۰، D= ۰.۱ گیگابایت:

$$M = ۱۰۰ \times ۰.۱ \times ۱.۵ = ۱۵ \text{ گیگابایت}$$

- **حالت ۲: آموزش مدل(Training)**
  - **۲ =K_T**
  - **مثال** :اگر R = ۱۰۰، D= ۰.۲ گیگابایت:

$$M = ۱۰۰ \times ۰.۲ \times ۲ = ۴۰ \text{ گیگابایت}$$

- **حالت ۳: پردازش زبان طبیعی(NLP)**
  - **۲.۵ =K_T**
  - **مثال** :اگر R = ۱۰۰، D= ۰.۳ گیگابایت:

$$M = ۱۰۰ \times ۰.۳ \times ۲.۵ = ۷۵ \text{ گیگابایت}$$

# 4. پارامترهای تأثیرگذار دیگر

## 4.1.نوع مدل هوش مصنوعی

- **مدل‌های کوچک** :نیاز به GPU و RAM کمتر.
- **مدل‌های بزرگ) مانند GPT-3 یا :(BERT** نیاز به GPU و RAM بیشتر.

## 4.2. اندازه‌ی دسته‌ی پردازش(Batch Size)

- **Batch Sizeبزرگ** :نیاز به GPU و RAM بیشتر.
- **Batch Sizeکوچک** :نیاز به GPU و RAM کمتر.

## 4.3. نوعGPU

- **GPUهای قدرتمند) مانند :(NVIDIA A100** نیاز به تعداد GPU کمتر.
- **GPUهای معمولی) مانند :(NVIDIA T4** نیاز به تعداد GPU بیشتر.

---

# 5. مثال جامع

فرض کنید یک اپلیکیشن هوش مصنوعی دارید که:

- **تعداد درخواست‌های همزمانR** : ۲۰۰
- **نوع درخواست** :پردازش زبان طبیعی(NLP)
- **حجم داده هر درخواست(D)** : ۰.۲ گیگابایت
- **زمان پردازش هر درخواستP** : ۰.۲ ثانیه
- **ظرفیت پردازشی هر GPU بهC**: TFLOPS۱۰
- **ضریب استفاده از(U) GPU** : ۰.۸
- **ضریب پیچیدگی نوع درخواست(F_T)** : ۲

- **ضریب حافظه‌ی مورد نیاز :(K_T)** ۲.۵

## 5.1. محاسبه‌ی تعدادGPU ها

$$G = \left\lceil \frac{۲۰۰ \times ۰.۲ \times ۲}{۱۰ \times ۰.۸} \right\rceil = \left\lceil \frac{۸۰}{۸} \right\rceil = ۱۰$$

## 5.2. محاسبه‌ی میزانRAM

$$M = ۲۰۰ \times ۰.۲ \times ۲.۵ = ۱۰۰ \text{ گیگابایت}$$

$M$=۲۰۰×۰.۲×۲.۵=۱۰۰ گیگابایت

---

## 6.خلاصه این بخش

با استفاده از این فرمول‌ها و پارامترها، می‌توانید به‌راحتی تعدادGPU ها و میزان RAM مورد نیاز برای پردازش درخواست‌های هوش مصنوعی را محاسبه کنید. این رویکرد به شما کمک می‌کند تا بر اساس نیازهای خاص خود، منابع سخت‌افزاری را به‌صورت بهینه‌تری تخصیص دهید.

شرکت صبا (دیجی هوش هما)

## راه های تماس با ما


| | |
|---|---|
| ایمیل | cmo@sabaind.com |
| تلگرام | @DigiHooshHoma |
| لینکداین | https://www.linkedin.com/company/maaia/ |
| وب سایت | www.sabaind.com |
| موبایل | +98-0905 611 0895 |
| تلفن | +98-021-22146343 |
| آدرس | تهران – سعادت آباد – میدان بهرود – خیابان عابدی – ساختمان صبا |

# برای طراحی دیتاسنتر خود با ما تماس بگیرید

- ماموریت تیم دیجی هوش هما، آشناسازی اعضای محترم هیات مدیره، مدیران عامل، مدیران ارشد و متخصصین سازمانها برای چگونگی پیاده‌سازی و استفاده از فناوریهای هوشمندسازی، و به ویژه هوش‌مصنوعی به عنوان فناوری شالوده شکن عصر حاضر، در سازمان یا دپارتمان مدنظر آنها می‌باشد.
- ما باور داریم که امروزه تحول دیجیتال و پیاده‌سازی هوش‌مصنوعی در سازمان‌ها نه یک انتخاب، بلکه یک ضرورت است، چرا که هزینه عدم استفاده از آن برای شرکت زیان بار خواهد بود. ما در مجموعه دیجی هوش هما (صبا) با افتخار آماده‌ایم تا شما را در این مسیر همراهی کنیم.
- مجموعه دیجی هوش هما (صبا) در نقش مشاوره، آموزش، و ارائه راه‌حلهای هوش‌مصنوعی به شما و سازمان شما متعهد بوده و پیاده‌سازی بهینه راه‌حلهای هوش‌مصنوعی را با کمک برترین متخصصین و شرکتهای ارائه دهنده خدمات هوش‌مصنوعی برای شما رقم خواهد زد.
- تکنولوژی‌های هوشمندسازی و هوش‌مصنوعی می‌توانند تغییرات اساسی در نحوه کارکرد و بهره‌وری سازمان‌ها به وجود آورند. با استفاده از راه‌حل‌های ما، می‌توانید فرایندهای پیچیده را ساده‌سازی کنید تصمیم‌گیری‌های بهتری انجام دهید و بهره‌وری سازمان خود را به شکل چشمگیری افزایش دهید.
- چشم‌انداز ما در دیجی هوش هما (در نقش اپراتور هوش‌مصنوعی AI Integrator)، ایجاد تحولی واقعی و ملموس در سازمان‌های متقاضی است. ما با تیمی‌مجرب و متخصص، آماده‌ایم تا بهترین مشاوره‌ها و راه‌حل‌های هوشمند را به شما ارائه دهیم. از مشاوره، آموزش‌های درون‌سازمانی تا پیاده‌سازی کامل راه‌حل‌های هوش‌مصنوعی و اینترنت اشیا، ما در تمامی‌مراحل همراه شما هستیم.
- منتظر تماس شما متقاضی گرامی هستیم تا با همدیگر آینده‌ای هوشمندتر و پر رونق‌تر برای مملکت خود بسازیم. برای دریافت اطلاعات بیشتر و مشاوره تخصصی، با ما تماس بگیرید.