# Case Study 4: Accelerated Computing Instances

## 1. Overview of Recommended Instances

Accelerated Computing instances are optimized for workloads that need **hardware acceleration**—mainly using **GPUs** or **specialized chips** for tasks like deep learning inference, rendering, and high-performance computing (HPC).

Recommended instance types—**P4, P5, G5, and G6**—are ideal for **AI/ML workloads**, **3D rendering**, and **video encoding**, where traditional CPUs would be too slow or inefficient.

- **P-Series (P4/P5)**: Built for **AI training and inference**, offering high-performance **NVIDIA GPUs** like A100 (P4) and H100 (P5).
- **G-Series (G5/G6)**: Targeted at **graphics-intensive applications** like gaming, remote workstations, and media processing.

## 2. Characteristics

| Instance Series | Key Characteristics |
|---|---|
| **P4** | Uses NVIDIA A100 GPUs; optimized for deep learning inference and training. |
| **P5** | Newer than P4, uses NVIDIA H100 GPUs; ideal for large-scale ML models and HPC. |
| **G5** | NVIDIA A10G GPUs; great for game streaming, ML inference, and graphics workloads. |
| **G6** | Upcoming Graviton-powered GPU instances; cost-effective and optimized for modern graphics workloads. |

## 3. Why They Are Suitable

### Scenario A (3D NPC Animation System)

Needs **fast inference** and **real-time rendering** for AI-powered character movements.

- **P4/P5** provide high-throughput GPU compute for deep learning model inference.
- **G5** is well-suited for **3D rendering with GPU** and **ML-powered NPC behaviour generation**.

### Scenario B (Video Rendering and Encoding)

Requires fast, parallel processing of large media files.

- **G5/G6** are optimized for **video encoding**, **remote rendering**, and **graphics-intensive processing**.
- **P4** may also be used for ML-enhanced video effects or filters.

## 4. Consideration Detailing the Instance Series

| Series | GPU Type | Memory Bandwidth | Focus Area |
|---|---|---|---|
| P4 | NVIDIA A100 | High | AI/ML inference & training |
| P5 | NVIDIA H100 | Very High | Large-scale deep learning, HPC |
| G5 | NVIDIA A10G | Medium-High | Graphics, ML inference, video processing |
| G6 | Graviton + GPU | TBD | Graphics & cost-optimized workloads |

## 5. Comparison and Selection

| Use Case | P4 | P5 | G5 | G6 |
|---|---|---|---|---|
| GPU Power | Better | Best | Ok | Ok |
| Deep Learning Inference | Best | Best | Ok | Ok |
| 3D Character Animation (ML + GPU) | Best | Best | Better | Ok |
| Video Encoding & Rendering | Better | Better | Best | Better |
| Cost Efficiency | Ok | No | Better | Best |
| Graphics/Gaming Optimization | No | No | Best | Best |
| Future-ready/Cloud-native | Ok | Best | Better | Best |

## 6. Key Considerations Supporting the Business Case

- **AI Acceleration**:
  P-Series (especially **P5**) are tailored for **deep learning inference**, which is core to the NPC engine's intelligence.
- **Rendering & Encoding**:
  **G5/G6** offer **optimized GPU rendering**, perfect for media and gaming workloads with real-time processing needs.
- **Scalability**:
  These instances support **Elastic Fabric Adapter (EFA)** for low-latency multi-node communication (especially P-series), great for scaling AI workloads.
- **Cost vs Performance**:
  G6 is expected to offer **cost-effective performance** for **mid-level GPU tasks**, while **P5** is suitable for enterprise-grade ML workloads where speed trumps cost.

## 7. Conclusion

- For **AI-powered 3D animation and real-time NPC behavior**, **P4/P5** are the best choices with **fast GPU inference**.
- For **video encoding and rendering**, **G5** provides the right mix of **graphics power and cost efficiency**.
- Choose **G6** for **modern, cost-effective GPU workloads** especially if transitioning to **Graviton-powered architectures**.
- Selection should consider **performance needs**, **GPU type**, **cost sensitivity**, and **future scalability**.