Duy trì thêm dung lượng dự phòng để xử lý các đột biến.

Cấu hình thời gian chờ (cooldown) để tránh việc scale lên/xuống liên tục.

Hạn chế: có thể làm tăng chi phí do tài nguyên nhàn rỗi.

3. Chiến lược quản lý

Dự báo năng lực (Predictive Capacity Planning)

Dựa vào lịch thi để ước lượng tải cao.

Đặt trước dung lượng tài nguyên đám mây.

Kịch bản khẩn cấp (Emergency Playbooks)

Bộ quy trình sẵn có cho việc mở rộng hoặc khởi động lại dịch vụ.

Suy giảm có kiểm soát (Graceful degradation): giảm chất lượng video hoặc vô hiệu hóa các tính năng không thiết yếu khi quá tải.

Ưu tiên tính năng (Feature Prioritization)

Dịch vụ quan trọng (nộp bài, quiz, đăng nhập) phải luôn duy trì hoạt động.

Dịch vụ phụ (AI gợi ý, phân tích) có thể tạm ngưng khi cần.

4. Lý do

Khớp tài nguyên với nhu cầu: chỉ mở rộng khi cần để tiết kiệm chi phí.

Đảm bảo tốc độ & tính sẵn sàng: sinh viên phải luôn truy cập ổn định.

Khả năng chịu lỗi & bền vững: triển khai đa vùng, giám sát, và dung lượng dự phòng giúp hệ thống liên tục hoạt động.

5. Đánh đổi

Chi phí ↔ Hiệu năng: Nhiều tài nguyên đảm bảo mượt mà nhưng chi phí cao. Autoscaling giảm chi phí nhưng có thể trễ vài giây.

Đơn giản ↔ Linh hoạt: Dịch vụ đám mây quản lý sẵn thì dễ dùng nhưng dễ bị khóa nhà cung cấp; giải pháp tùy chỉnh linh hoạt hơn nhưng đòi hỏi chuyên môn cao.

Độ trễ ↔ Bảo mật: CDN/bộ nhớ đệm tăng tốc nhưng cần bảo vệ dữ liệu nhạy cảm.

Độ phức tạp ↔ Độ tin cậy: Nhiều kỹ thuật (multi-region, predictive scaling) tăng độ bền vững nhưng làm hệ thống phức tạp hơn.