

AI ENGINEERING

AzkenOS

Proyecto: Ignacio Bonilla





Problema:

- En nuestra empresa contamos con un departamento técnico encargado de la producción de equipos de alto rendimiento: estaciones de trabajo HPC, servidores de cómputo y sistemas optimizados para distintos sectores. Principalmente, trabajamos con servidores que integran entre 2 y 10 GPUs NVIDIA de diversas gamas y generaciones.
- Actualmente, los compañeros del área de producción deben realizar manualmente los benchmarks y validar cada máquina de forma individual. Esto implica ejecutar todos los comandos a mano, revisar los resultados uno por uno y generar informes con capturas de pantalla que luego deben ser editados y mejorados antes de su entrega.
- Este proceso, además de ser muy tedioso y repetitivo, incrementa el tiempo de validación y producción, y también el riesgo de errores humanos durante la ejecución de pruebas o la documentación de los resultados.

Solución

- He desarrollado un **Live de Ubuntu personalizado** con **Ollama** y los **drivers de NVIDIA** ya integrados además de toda una suite de Benchmarks, capaz de arrancar sin instalación. Este entorno ejecuta el modelo **QWEN2.5**, seleccionado por su excelente rendimiento **tanto en GPU como en CPU**, lo que lo hace ideal para todos los equipos que montamos, incluso aquellos sin tarjeta gráfica dedicada.

Automatización según el tipo de equipo:

-  **GPU + BMC** → **Extrae temperaturas, voltajes y frecuencias vía ipmitool.**
-  **GPU sin BMC** → **Usa sensors y herramientas del sistema para obtener los datos clave.**
-  **Sin GPU + BMC** → **Monitorización térmica y de voltaje usando la controladora BMC.**
-  **Sin GPU ni BMC** → **Obtiene toda la información desde sensors del sistema.**

Resultado final

- Todos los **benchmarks** generan archivos **log** que el **modelo IA analiza automáticamente**. A partir de los resultados, genera un **informe en Discord**, creando un **hilo con el número de serie del equipo**:
- Si detecta problemas, **menciona al administrador técnico**.
- Si todo está correcto, **genera el informe sin alertas**, optimizando tiempos y reduciendo errores humanos.



Inicio del sistema operativo

GNU GRUB version 2.12

```
*? Iniciar Azken Muga OS (modo normal)
? Iniciar Azken Muga OS (modo gráfico seguro)
? Instalación OEM personalizada por Azken Muga
? Visitar web de Azken Muga
? Ajustes de Firmware UEFI
? Arrancar desde el siguiente dispositivo
```

Use the ▲ and ▼ keys to select which entry is highlighted.
Press enter to boot the selected OS, `e' to edit the commands
before booting or `c' for a command-line.
The highlighted entry will be executed automatically in 20s.

Despliegue

Azken-IA **APP** 22:33
 **Automatizado iniciado** AgentOS-Keepcoding
 <http://192.168.1.3:5000/>

```
Could not open device at /dev/ipmi0 or /dev/ipmi/0 or /dev/ipmidev/0: No such file or directory
Could not open device at /dev/ipmi0 or /dev/ipmi/0 or /dev/ipmidev/0: No such file or directory
[0;36mAzken Muga 🌟💎[0m
```

```
[0;32mSistema optimizado para pruebas de benchmarks.[0m
```

Estadísticas del sistema:

- GPUs NVIDIA CUDA: 1 (NVIDIA GeForce RTX 3090,)
- CPUs disponibles: 8
- Modelo de CPU: Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz

BIOS Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz To Be Filled By O.E.M. CPU @ 4.5GHz

- Frecuencia actual de CPU: 3700.40 MHz
- Frecuencia base (mínima): 800
- Frecuencia boost (máxima): MHz 4.90 GHz
- Núcleos físicos: 8
- Hilos por núcleo: 1
- Caché L1d: 256 KiB
- Caché L1i: 256 KiB
- Caché L2: 2 MiB
- Caché L3: 12 MiB
- Memoria RAM total: 16G

Información del BMC:

- IP del BMC: [0;31mNo disponible[0m
- MAC del BMC: [0;31mNo disponible[0m
- Placa base: Z390 UD

```
[0;36m🐞 Azken Muga LABS [0m
```

Interfaz de usuario

← → ↻ No es seguro 192.168.1.3:5000

Personal Estudios Varios Correos

Todos los marcadores

Archivos en /

Seleccionar archivo | Ningún archivo seleccionado [Subir archivo](#)

Nombre	Acciones	Estado
tests/	Descargar carpeta	-
tools/	Descargar carpeta	-
adaptdscript.sh	Editar Interactuar Descargar	-
AgentOS-Keepcoding.txt	Editar Descargar	-
automatizadodesktop.sh	Editar Interactuar Descargar	-
automatizadoserver.sh	Editar Interactuar Descargar	-
discordagent.py	Editar Interactuar Descargar	-
individual.sh	Editar Interactuar Descargar	-
limpieza.sh	Editar Interactuar Descargar	-
seleccion.py	Editar Interactuar Defensar Descargar	En ejecución
singpudesktop.sh	Editar Interactuar Descargar	-
singpuserver.sh	Editar Interactuar Descargar	-
stream	Editar Descargar	-
valores.sh	Editar Interactuar Descargar	-
webaplicacion.py	Editar Interactuar Descargar	-

Métricas

CPU Uso: 3.5%
CPU Temp: 53°C
CPU Freq: 2297.2 MHz
RAM Uso: 20.4%
GPU0 (NVIDIA GeForce RTX 3090):
Carga: 0% Mem: 1.5% Temp: 51°C Freq: 210 MHz

Ejecución de benchmarks

```
← → ↻ ⚠ No es seguro 192.168.1.3:5000/terminal/seleccion.py ☆ 🌐 📁 📌 🔍
Personal Estudios Varios Correos ☰
Levantando 'ollama serve' en segundo plano...
time=2025-07-21T22:39:57.099+02:00 level=INFO source=routes.go:1235 msg="server config" env="map[CUDA_VISIBLE_DEVICES: GPU_DEVICE_ORDINAL: HIP_VISIBLE_DEVICES: HSA_OVERRIDE_GFX_VERSION: HTTPS_PROXY: HTTP_PROXY: NO_PROXY: OLLAMA_CONTEXT_LENGTH:4096 OLLAMA_DEBUG:INFO OLLAMA_FLASH_ATTENTION:false OLLAMA_GPU_OVERHEAD:0 OLLAMA_HOST:http://127.0.0.1:11434 OLLAMA_INTEL_GPU:false OLLAMA_KEEP_ALIVE:5m0s OLLAMA_KV_CACHE_TYPE:OLLAMA_LLM_LIBRARY: OLLAMA_LOAD_TIMEOUT:5m0s OLLAMA_MAX_LOADED_MODELS:0 OLLAMA_MAX_QUEUE:512 OLLAMA_MODELS:/root/.ollama/models OLLAMA_MULTIUSER_CACHE:false OLLAMA_NEW_ENGINE:false OLLAMA_NOHISTORY:false OLLAMA_NO_PRUNE:false OLLAMA_NUM_PARALLEL:0 OLLAMA_ORIGINS:[http://localhost https://localhost http://localhost:* https://localhost:* http://127.0.0.1 https://127.0.0.1 http://127.0.0.1:* https://127.0.0.1:* http://0.0.0.0 https://0.0.0.0 http://0.0.0.0:* https://0.0.0.0:* app://* file://* tauri://* vscode-webview://* vscode-file://*] OLLAMA_SCHED_SPREAD:false ROCR_VISIBLE_DEVICES: http_proxy: https_proxy: no_proxy:]"
time=2025-07-21T22:39:57.101+02:00 level=INFO source=images.go:476 msg="total blobs: 5"
time=2025-07-21T22:39:57.101+02:00 level=INFO source=images.go:483 msg="total unused blobs removed: 0"
time=2025-07-21T22:39:57.101+02:00 level=INFO source=routes.go:1288 msg="Listening on 127.0.0.1:11434 (version 0.9.3)"
time=2025-07-21T22:39:57.108+02:00 level=INFO source=gpu.go:217 msg="looking for compatible GPUs"
time=2025-07-21T22:39:57.530+02:00 level=INFO source=types.go:130 msg="inference compute" id=GPU-222d791d-cff0-9408-d395-affd2b50d977 library=cuda variant=v12 compute=8.6 driver=12.9 name="NVIDIA GeForce RTX 3090" total="23.6 GiB" available="23.2 GiB"
Ejecutando 'ollama pull gwen2.5'...
[GIN] 2025/07/21 - 22:40:01 | 200 | 4.971942ms | 127.0.0.1 | HEAD | "/"
[GIN] 2025/07/21 - 22:40:02 | 200 | 648.565183ms | 127.0.0.1 | POST | "/api/pull"

Seleccione el script que desea ejecutar:
1. Automatizado Server
2. Automatizado Desktop
3. Sin Gpu Server
4. Sin Gpu Desktop
5. Individual
Ingrese su opción: █
```

Benchmark #1

MONITORIZACION EN TIEMPO REAL



RESULTADOS DE LOS BENCHMARKS



ANALISIS DE QWEN2.5 DE LOGS

```
Azken-IA APP 22:48

Monitorización en tiempo real para AgentOS-Keepcoding

Hora: 22:42:05

• CPU: 0.4%
• RAM: 27.5%
• Temperatura CPU: 22.3°C
• GPU: 100.0% / 68.0°C
• Frecuencia CPU: 3795.0 MHz
• Uso de Disco: 20.9%
• Red: +0.01 MB enviados / +0.01 MB recibidos
• Carga del Sistema (1 min): 0.58

Últimas 5 Líneas (AgentOS-Keepcoding.txt)
D[0:36m\ Azken Muga LABS D[0m
Ejecutando Gobernador...
Cambiando el gobernador de la CPU a 'performance'
para todas las 8 CPUs.
Todas las CPUs han sido configuradas en el modo
'performance'.
Ejecutando Octane...
(editado)
```

Nos permite ver las últimas 5 líneas del log que se van generando en tiempo real.

```
Azken-IA APP 23:12

Se van a ejecutar los siguientes scripts en secuencia con
1. Gobernador
2. Octane
3. Geekbench GPU
4. AnchoBanda
5. MprimeDesktop -> num_cpus: 1, mem_reservada_gb: 6, fre

Expandir  AgentOS-Keepcoding_final.txt 18 KB

AgentOS-Keepcoding_reports.zip
58.38 KB

/
system_info.txt (14.83 KB)
journalctl.txt (281.03 KB)
bios.txt (25.36 KB)
results_0.txt (14.16 KB)
cpu_clocks.txt (7.50 KB)
resultado.html (12.30 KB)
informe.html (17.06 KB)
```

Sube los documentos generados durante las pruebas y el log sin tratamiento de la IA

Análisis AgentOS-Keepcoding
Azken-IA: Hiló de análisis generado automáticamente.
6 · hace 1 min

informe-agente-ia
Análisis AgentOS-Keepcoding

AgentOS-Keepcoding_informe.pdf
34.14 KB

Lee los logs y genera un informe, mencionando al administrador cuando haya que revisar una parte del log por sospecha de fallo

Benchmark #2

- La IA detecta un error, he probado una GPU RTX 3090 que está en mal estado para este benchmark, el speedclock base de una RTX 3090 es 1395 Mhz por lo cual está por debajo de su base y nuestro agente lo ha podido identificar en los logs.

Mención del agente al administrador en discord

@Akantor posible incidencia en GPU-BURN:

Análisis del Test GPU-BURN

Resumen General:

El test "GPU-BURN" se realizó para evaluar la estabilidad térmica y de rendimiento de una NVIDIA GeForce RTX 3090 durante un periodo de 300 segundos.

Detección de Mensajes de Error:

No se observaron mensajes directos de error en la sección de logs proporcionada. Sin embargo, los siguientes puntos sugieren posibles problemas:

1. Función por debajo del clock base configurado:

- Se reporta que el promedio de MHz para la GPU 0 es de 1252.5510 MHz, lo cual representa un 11.00% por debajo del clock base configurado.

Parte del LOG donde está el error.

```
Killing processes with SIGTERM (soft kill)
Tiempo transcurrido: 00:04:52 | Tiempo restante: 00:00:08
Tiempo transcurrido: 00:04:54 | Tiempo restante: 00:00:06
Using compare file: compare.ptx
Burning for 300 seconds.
Initialized device 0 with 24123 MB of memory (23750 MB available, using 21375 MB of it), using FLOATS, using Ten
Results are 268435456 bytes each, thus performing 81 iterations
Freed memory for dev 0
Unmited cublas
Tiempo transcurrido: 00:04:56 | Tiempo restante: 00:00:04
Tiempo transcurrido: 00:04:58 | Tiempo restante: 00:00:02
Deteniendo el proceso de GPU-BURN...
Proceso de GPU-BURN detenido.
Calculando promedios de resultados...

GPU - 0: Suma total de MHz: 184125 MHz
GPU - 0: Líneas extraídas: 147
GPU - 0: Promedio de MHz (100 %): 1252.5510 MHz
GPU - 0: Funciona por debajo de la base por: 11.00 %
GPU - 0: Líneas que no cumplen con el criterio: 1
```

Configuración donde el técnico de producción indica el speedclock de esa GPU

Se van a ejecutar los siguientes scripts en secuencia con los siguientes parámetros:

1. Gobernador
2. Octane
3. Geekbench GPU
4. AnchoBanda
5. MprimeDesktop -> num_cpus: 1, mem_reservada_gb: 6, frecuencia_base_mhz: 3600, duración: 5 minutos
6. Geekbench
7. GPU-BURN -> clock base: 1395 MHz, duración: 5 minutos
8. FIO
9. Sectores
10. Resultado
11. Comprobaciones -> Dispositivo PCI: Intel