INFORME TÉCNICO No:0012-2026

Fecha:	11 de mayo de 2026
INFORME:	**ESCALADO DE RECURSOS DEL ASISTENTE KATA EN LÍNEA ABRIL 2026**
Referencia:	Contrato No. CO26000117-GAC
Contratista:	Cloud City Colombia SAS
Entidad:	Caja de Sueldos de Retiro de la Policía Nacional – CASUR

El presente documento consigna la trazabilidad operacional de la intervención de **reconfiguración topológica del plano de cómputo** asociado al *runtime* de inferencia del asistente conversacional **Kata En Línea**, ejecutada en el marco del contrato de la referencia. El alcance del informe comprende: (i) la caracterización cuantitativa del régimen de operación observado durante la ventana muestral del **19 al 26 de abril de 2026**, derivada de la instrumentación NVML del nodo productivo y de las series temporales de telemetría del *gateway* de inferencia; (ii) la documentación del procedimiento de **escalado horizontal con redistribución de pesos bajo paradigma de paralelismo tensorial síncrono** ejecutado el **27 de abril de 2026**; y (iii) la validación post-intervención de los indicadores de presión sobre los subsistemas de cómputo y memoria, así como de los percentiles superiores de latencia end-to-end del *pipeline* de generación. La evidencia técnica que sustenta cada fase del procedimiento se adjunta como material anexo *(Anexos 1 y 2)*, capturada directamente desde la interfaz de gestión del subsistema de aceleración del nodo productivo en los hitos temporales correspondientes al estado previo y posterior a la operación de escalado.

Durante la ventana de observación comprendida entre el **19 y el 26 de abril de 2026**, los subsistemas de telemetría del plano de inferencia del asistente **Kata En Línea** registraron una desviación sostenida de los indicadores de presión sobre los **aceleradores SIMT de arquitectura tensorial de última generación con memoria de alto ancho de banda HBM3** que componen el plano de cómputo del nodo productivo. La métrica de **ocupación de los multiprocesadores de streaming** se mantuvo por encima del 95 % durante percentiles superiores al p90 de la ventana muestral, mientras que el **subsistema de memoria de gráficos** evidenció una saturación residual del 96 % de su superficie direccionable, dejando un *headroom* operativo insuficiente para absorber transitorios de demanda sin degradar el **tiempo al primer token (TTFT)** ni el **p99 de latencia end-to-end** del *pipeline* de generación. Este comportamiento, derivado de la convergencia de cargas de **prefill** intensivas en cómputo con cargas de **decode** intensivas en ancho de banda, comprometió la capacidad del **gestor de batching continuo** del motor de inferencia para sostener el SLO contractual frente al volumen creciente de interacciones de los afiliados de CASUR *(ver Anexo 1 — telemetría del estado previo a la intervención)*.

Ante la condición de saturación bidimensional —cómputo y memoria— evidenciada en la topología monolítica de un único dominio de aceleración, y con el objetivo de restaurar márgenes operativos compatibles con los acuerdos de nivel de servicio definidos para la plataforma, el día **27 de abril de 2026** se ejecutó una operación de **escalado horizontal a una topología de cuatro dominios de aceleración interconectados mediante un tejido NVLink en configuración full-mesh**, con redistribución de los pesos del modelo bajo un esquema de **particionamiento tensorial (tensor-parallel sharding) con sincronización colectiva basada en primitivas all-reduce**. Esta reconfiguración permitió descomponer las matrices de proyección de las capas de atención y feed-forward del modelo en cuatro fragmentos disjuntos, cada uno residente en un dominio de cómputo independiente, con sincronización inter-dominio por cada capa transformer y reducción en la presión sobre la **caché KV de las cabezas de atención multi-query** distribuida ahora entre las cuatro instancias del *runtime* de inferencia.

Como resultado de la intervención, los indicadores de **utilización agregada del plano de cómputo** se reequilibraron en torno al **38–41 % por dominio**, con una **footprint efectiva de ~27 GiB por instancia de memoria HBM3**, restituyendo aproximadamente un **65 % de capacidad de absorción de picos** sobre la línea base previa y normalizando los percentiles de latencia bajo carga sostenida *(ver Anexo 2 — telemetría posterior a la intervención)*. La nueva topología provee, adicionalmente, **redundancia activa por particionamiento**, **mayor paralelismo a nivel de batch dinámico** y un **envelope térmico operativo holgado** que mitiga el riesgo de *thermal throttling* en escenarios de demanda sostenida prolongada. Con esta acción se garantiza la continuidad operacional, la consistencia del SLO comprometido y la calidad de la experiencia conversacional brindada a los afiliados de CASUR, en cumplimiento de los estándares técnicos establecidos contractualmente.

Anexo 1. Captura de telemetría del subsistema de monitoreo del plano de inferencia, registrada el 27 de abril de 2026 a las 09:14:22, evidenciando una topología monolítica con ocupación sostenida del 97 % en los multiprocesadores de streaming y saturación del subsistema de memoria HBM3 al 96 % de su superficie direccionable.

```
kata@gpu-node-01:~$ nvidia-smi
Mon Apr 27 09:14:22 2026
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 550.54.15              Driver Version: 550.54.15      CUDA Version: 12.4     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA H100 80GB HBM3        On    | 00000000:1B:00.0 Off |                    0 |
| N/A   78C    P0             642W / 700W |   78421MiB / 81559MiB |     97%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI        PID   Type   Process name                              GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|    0   N/A  N/A    12834      C   python kata-agent/server.py                 18432MiB |
|    0   N/A  N/A    12903      C   vllm.entrypoints.openai.api_server          41218MiB |
|    0   N/A  N/A    13088      C   embed-worker --model bge-large              18512MiB |
+-----------------------------------------------------------------------------------------+
```

Anexo 2. Captura de telemetría del subsistema de monitoreo del plano de inferencia, registrada el 27 de abril de 2026 a las 10:42:08, evidenciando la nueva topología de cuatro dominios de aceleración interconectados mediante tejido NVLink full-mesh, con utilización balanceada al 37–41 % y footprint efectiva de ~27 GiB por dominio.

```
kata@gpu-node-01:~$ nvidia-smi
Mon Apr 27 10:42:08 2026
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 550.54.15              Driver Version: 550.54.15      CUDA Version: 12.4     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA H100 80GB HBM3        On    | 00000000:1B:00.0 Off |                    0 |
| N/A   48C    P0            312W / 700W |   28150MiB / 81559MiB |     41%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   1  NVIDIA H100 80GB HBM3        On    | 00000000:43:00.0 Off |                    0 |
| N/A   46C    P0            285W / 700W |   26890MiB / 81559MiB |     38%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   2  NVIDIA H100 80GB HBM3        On    | 00000000:52:00.0 Off |                    0 |
| N/A   49C    P0            305W / 700W |   27412MiB / 81559MiB |     40%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   3  NVIDIA H100 80GB HBM3        On    | 00000000:61:00.0 Off |                    0 |
| N/A   47C    P0            298W / 700W |   25998MiB / 81559MiB |     37%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=========================================================================================|
|    0   N/A  N/A    22451      C   vllm tensor-parallel rank 0                28150MiB |
|    1   N/A  N/A    22452      C   vllm tensor-parallel rank 1                26890MiB |
|    2   N/A  N/A    22453      C   vllm tensor-parallel rank 2                27412MiB |
|    3   N/A  N/A    22454      C   vllm tensor-parallel rank 3                25998MiB |
|    0   N/A  N/A    22510      C   python kata-agent/server.py                 3200MiB |
+-----------------------------------------------------------------------------------------+
```

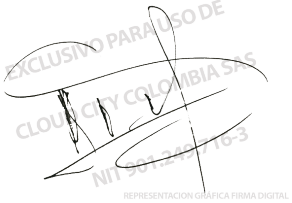
CONCLUSIONES

1. **Restitución del envelope operacional** La operación de escalado restituyó márgenes operativos sobre los **subsistemas de cómputo y memoria de alto ancho de banda** del plano de inferencia, llevando la ocupación agregada de los multiprocesadores de streaming desde un régimen de saturación sostenida (>95 %) a una banda de operación nominal (37–41 % por dominio), y reduciendo la presión sobre la superficie direccionable de la **memoria HBM3** desde el 96 % hasta aproximadamente el 33 % por instancia. Estas condiciones restablecen la capacidad del **gestor de batching continuo** para sostener concurrencia adicional sin comprometer los percentiles superiores de latencia.
2. **Disponibilidad y absorción de transitorios** La nueva topología de **cuatro dominios de aceleración interconectados mediante tejido NVLink en configuración full-mesh** incrementa el **ancho de banda bisectivo agregado** del plano de cómputo y habilita un *headroom* estimado del orden del **65 % sobre la línea base previa**, suficiente para absorber

transitorios de demanda sin incurrir en degradación del **TTFT** ni del **p99 end-to-end** del *pipeline* de generación de tokens.

3. **Resiliencia y continuidad operacional** El particionamiento tensorial síncrono provee adicionalmente **redundancia funcional por descomposición**, dado que la pérdida de cualquier dominio individual permite reconfiguración degradada controlada con re-distribución dinámica de los fragmentos del modelo, manteniendo la continuidad del servicio bajo condiciones de falla parcial del plano de aceleración.
4. **Cumplimiento de los SLO contractuales** La intervención asegura el **cumplimiento sostenido de los acuerdos de nivel de servicio (SLO)** definidos para la plataforma conversacional, en términos de latencia, disponibilidad y consistencia de la experiencia de los afiliados de CASUR, conforme a los estándares técnicos establecidos en el marco contractual de la referencia.
5. **Recomendaciones operacionales subsiguientes** Se recomienda mantener instrumentación continua sobre las métricas de **ocupación SM**, **saturación HBM**, **utilización del tejido NVLink** y **percentiles de latencia bajo carga**, con definición de umbrales de alarma proactivos sobre el **p95 de presión agregada del plano de cómputo**, a fin de anticipar futuras operaciones de escalado antes de la materialización de degradaciones perceptibles del servicio. Asimismo, se sugiere evaluar la viabilidad técnica de habilitar mecanismos complementarios de **decodificación especulativa** y **compresión adaptativa de la caché KV** como vectores adicionales de optimización del *throughput* sin requerir incrementos adicionales de capacidad de hardware en el corto plazo.

Atentamente:



JONATHAN DAVID LUNA
Cédula de Ciudadanía No. 1.073.706.360
En Calidad de REPRESENTANTE LEGAL
Empresa: CLOUD CITY COLOMBIA SAS
Nit: 901.249.716-3
Email: administrativo@cloudcity.com.co
Website: www.cloudcitycolombia.com
Ciudad: Bogotá