SOFWERX

Hardware Enabled AI Acceleration Assessment Event (AE)
Q&A Telecon Transcript
04 DECEMBER 2025

1. **Has the Air Force evaluated deterministic GPU compute yielding bit-for-bit identical outputs on repeated runs? Are there baselines or policy constraints?**
   Could not answer on behalf of the Air Force.

2. **Are there edge, airborne, or ground environments where deterministic GPU behavior is especially valuable or required for certification?**
   This effort is not an edge/airborne/ground use case; it is an on-premises solution intended to support users at a single location.

3. **What level of traceability is required for mission-critical analytics—e.g., artifact hashes, signed manifests, hardware/driver provenance, or timestamped logs?**
   The effort is focused on hardware; This would be tied to accreditation/authorization requirements and would need follow-up with the appropriate stakeholders.

4. **Would a cryptographically verifiable chain—signed provenance + reproducible manifests—meet current audit, accreditation, or chain-of-custody standards?**
   The effort is focused on hardware; This would be tied to accreditation/authorization requirements and would need follow-up with the appropriate stakeholders.

5. **Is there a need to bundle GPU-accelerator software with the requested hardware?**
   The effort is focused on hardware; any software expectation is primarily limited to the basic GPU enablement stack (e.g., drivers).

6. **Will the AI vendors have to comply with the key AI supply chain security requirements and provisions from the most recent NDAA legislation?**
   We are not tracking further regulation criteria other than was specified in the assessment criteria.

7. **Can you provide additional details about the physical and environmental constraints of the remote site to ensure optimal system compatibility?**
   The remote site physical and environmental constraints would be provided in a written response after confirmation with the onsite stakeholders.

8. **Will the deployment site support 20kW per rack, or should we design for a lower power envelope?**
   Assume the solution will be less than 20kw per rack but will accept proposals with multiple tiers/options.

9. **For dual-network deployment, are there preferred methods or standards for enforcing physical and logical isolation?**
   Isolation would be enforced by keeping hardware single-homed to one network. If both unclassified and secret deployments are needed, separate hardware sets would be used per network.

10. **Does the target solution have any ruggedization requirements (extended temperature, shock, vibration, humidity, etc.) or SWaP-constraints?**
    No ruggedization or transportability constraints are expected; the system will be installed in a server room with adequate cooling.

11. **Will the government provide specific LLMs or datasets for validation, or should we pre-validate compatibility with open-source and commercial models?**
    Vendors should pre-validate compatibility with open-source and commercially available models; no specific models or datasets will be provided for validation.

12. **Are there existing frameworks that the deployed system should integrate with, or should we provide a self-contained orchestration layer?**
    We have existing frameworks/orchestration platforms that we will manage internally. We just need the hardware delivered, racked, powered, networked, and running a base operating system that our team can configure.

13. **Will the AE include live workload testing or primarily technical and architectural evaluation? Are there specific benchmarks or scenarios we should prepare for?**
    The Assessment Event is expected to be primarily a technical and architectural evaluation, while remaining open to vendors providing workload testing as part of their approach.

14. **Will there be opportunities to scale the solution beyond the initial site or integrate into broader SOCOM infrastructure?**
    There may be opportunities to scale beyond the initial site and integrate more broadly depending on results, but no guarantees were provided.

15. **Is USSCOM looking for solutions that distribute smaller numbers across multiple deployed systems, or solutions that aggregate much larger quantities of GPUs?**
    Both distributed and aggregated GPU approaches are of interest in general, but this specific event is focused on a smaller system for a deployed location.

16. **Your description of the system desired indicates a requirement to scale. How much scaling needs to be incorporated into the configuration proposed?**
    While there may be an option to scale in the future, incorporation into this proposal is not required.

17. **Is the desired scaling within the same chassis/rack? Can you better define this requirement.**
    Assume solution can be installed in a single rack or up to a maximum of 3 different racks.

18. **For the Jan 6 -8 Assessment Event, do you expect the vendor to have the configured HW/SW suite available for demo?**
    Vendors may pitch, demonstrate, and/or discuss solutions during one-on-one sessions; demonstrating a configured solution is welcome but not mandatory.

19. **Can the demo portion of the assessment be done virtually, or does equipment need to be brought into the assessment?**
    The Assessment Event supports both virtual participation (e.g., Teams) and in-person options; any hardware shown onsite would need to be standalone and not connected to SOFWERX systems.

20. **Please elaborate on the software requirements for the system, as well as operational and sustainment support requirements.**
    There is no specific mission software requirement for delivery beyond what is needed to enable GPU use (e.g., drivers). The intent is for the government team to deploy and continually update its own inference/model-serving and workflow software. For sustainment, the solution should not require vendor intervention to continue operating; limited vendor support for a couple of days after delivery may be requested.

21. **What software must be pre-installed for “turnkey” delivery?**
    "Turnkey” refers to providing the bare-metal OS and GPU drivers/enablement needed for the government to deploy its own inference and model-serving software, rather than including an MLOps platform or specific LLM frameworks.

22. **What latency and response time targets do you expect for 100+ concurrent users?**
    There were no hard latency or response-time targets established yet for 100+ concurrent users, and welcomed vendor expertise in proposing achievable minimal latency.

23. **Is this inference only, or do you require fine-tuning or training capabilities?**
    There is intent to support a mix of use cases: primarily inference/serving for production workflows, with some capacity for development and potentially fine-tuning or training.

24. **How much total storage capacity is needed?**
    The effort is not intended to address a storage shortfall; existing storage and databases are already in place, and the need is GPU acceleration. Storage requirements were described as not a primary factor.

25. **What read/write throughput is required for the workload?**

No specific read/write throughput requirement is defined; vendors may propose trade-offs and options.

26. **Do you need support for embeddings or vector stores?**
    No. The effort is focused on hardware.

27. **Is it 30A @ 110V or 208V?**
    Could not verify. Assume between 200-240V

28. **Which is prioritized for AI workloads: model accuracy (density) or performance (latency)?**
    It was indicated performance—specifically latency—will be prioritized, while acknowledging both accuracy and performance matter.

29. **Will models require custom datasets for training prior to onsite finetuning?**
    No. This effort is focused on hardware, not software or model tuning/development.

30. **Will the AI workloads primarily assess dynamic/streaming data (caching solutions required) or static data (large and long-term storage)?**
    The hardware should support both dynamic/streaming and static-data use cases, without constraining the use case.

31. **How often will all 100+ be using the AI edge network simultaneously and for how long?**
    Plan for a sustained load over time, with 100+ users potentially using the capability concurrently.

32. **Is the remote site expecting a self-contained scalable solution, or is the goal to upgrade current servers for AI workloads? Augment vs. Replace.**
    The goal is augmentation: existing compute and databases already support many workloads, and the intent is to add GPU capability to accelerate GPU-dependent workflows.

33. **How often will models/software need updates?**
    This effort is focused on hardware. The only software updates expected are for GPU stack.

34. **Will the remote location for this effort be CONUS or OCONUS, and what specific security requirements will apply?**
    The remote location is OCONUS and the intent is to support both unclassified and secret environments.

35. **Would it be possible to increase the maximum page count to allow for a more comprehensive and detailed response?**
    The maximum page count should remain as-is.

36. **What are the overall power restrictions and the per-rack power limitations for this effort?**
Server room is currently air cooled. No specific requirements, but expectation is for solutions to be self-contained and able to be installed in a standard server rack(s).

37. **What sizing metrics should be considered for the ROM, including memory, CPU, storage, and GPU requirements?**
Sizing around a configurable server supporting roughly 2–6 GPUs, with sufficient CPU and memory to fully utilize the attached GPUs, and looked to industry expertise to propose appropriate configurations.

38. **Could you please clarify how a "very large model" is defined according to the assessment criteria?**
Did not define a specific parameter number for "very large model," noting generally that such models require GPUs for performant execution. This effort is primarily focused on hardware to support AI efforts, with the only software being what is needed to enable GPU use (e.g., drivers).

39. **Please confirm whether there are additional workloads that need to be hosted on the solution beyond those requiring GPU—for example, middleware, databases, etc?**
No additional non-GPU workloads are intended to be hosted on the new solution; the site already has substantial compute and databases, and the goal is to augment existing workflows with GPU acceleration.

40. **Will there be any additional Q&A opportunities beyond the scheduled virtual teleconference prior to the submission deadline?**
There will be no additional virtual Q&A teleconferences beyond the teleconference; the Assessment Event is invite-only.

41. **Is liquid-cooling acceptable, or does SOCOM prefer air-cooled solutions to simplify long-term sustainment and field maintenance?**
The site is air-cooled only, but if the liquid cooling is self-contained it will be considered.

42. **Will the government provide baseline config requirements for processing CUI to ensure conformity with NIST 800-171?**
The focus is Impact Level 2 and Impact Level 6; Controlled Unclassified Information (CUI) requirements were not part of the stated scope.

43. **How will SOCOM measure energy efficiency and system performance—tokens per watt, throughput per watt, end-to-end latency, or another standard metric?**
Energy efficiency is not a consideration at this time. Please see the assessment criteria for this event for the comprehensive list.

44. **Should the deployed system integrate with existing orchestration, monitoring, or containerization frameworks, or provide a self-contained orchestration layer?**
    The system only needs to provide software to enable use of the GPUs (such as driver software).

45. **Can the government provide direction on fine-tuning use cases (e.g. model parameter sizes)?**
    It would not provide direction on fine-tuning use cases or parameter sizes; the focus is enabling use of popular open-source models, with fine-tuning as a secondary consideration.

46. **For the RAG use cases, are there estimates of context token requirements?**
    There are no finite context-token estimates; the software and workflows are expected to vary over time.

47. **Is SOCOM open to solutions that incorporate third-party software partners (RAG engines, monitoring tools, etc.) as part of a fully integrated capability?**
    Not required for this effort.

48. **Is the government interested in future support for non-GPU accelerators (ASIC, FPGA, NPUs), even if not required for Phase 1?**
    No.

49. **What is the power, cooling, and rack-space constraints at the intended deployment location(s), particularly for remote or austere environments?**
    The current site configuration needs to be confirmed. The current facility has a climate-controlled server room and enough power to moderately increase capacity with gpu enabled servers.

50. **What is the TOPS target with the system? and would better portability be desirable trade-off for moderately reduced processing?**
    There is no TOPS target at this time.

51. **What is the size in GB or TB of models you want to execute? - What's the size of the input data set and output data in TB you want to store?**
    Storage sizing is not the focus of this effort and did not provide model sizes in GB/TB; the problem being solved is insufficient for GPU capability rather than storage.

52. **Is the 30A rack capacity inclusive the entire rack, or available to each node.**
    Assume 30A for the entire rack but there is a possibility of having more than 1 rack and/or a dedicated rack.

53. **Is the mission application more data-driven (e.g., a very large LLM) or more computing-driven (e.g., smaller data, but intense dependencies) or both**
    The mission applications and corresponding software vary.

54. **Has the mission application already been deployed with an LLM/RAG or other AI/ML technologies?**
Yes, in a limited capacity which is currently testing current production hardware.

55. **Is redundant power required?**
Independent/redundant power not needed.

56. **Will SOFWERX provide a list of interested parties to facilitate/explore partnership opportunities (Small Business participation, etc.)?**
The teleconference chat could be used for introductions to facilitate partnerships, but no additional collaboration mechanism beyond that was indicated.

57. **Can you define percentage of activity in each workload such 70/30 A vs B, etc.**
Not determined yet.

58. **Is the 30A power capacity based on 120V (2.8kW) or 208V (4.9kW)? Should our proposed power consumption adhere to the 80% continuous capacity standard?**
Could not verify. Assume between 200V-240V

59. **Is standard data center rack space available (e.g., 42U rack, 1200mm deep), or are there specific constraints on server unit dimensions?**
Standard data rack

60. **What is the approximate size (in parameters) of the largest foundational LLM expected to load into GPU memory for inference and for fine-tuning?**
Did not specify a model size in parameters, stating a preference for running the largest models feasible within trade-offs among cost, power, and cooling.

61. **Are there specific Linux dists, containerization platforms or AI/ML frameworks that the solution must be compatible with, or can we propose the optimal stack?**
Responses may propose the optimal stack. The solution must provide servers which have all required software to enable immediate utilization of the hardware's GPUs.

62. **Have you down selected a large language model and RAG architecture? 2. Has a threshold and objective for tokens per second been established?**
No LLM or RAG architecture has been down-selected, and no threshold or objective for tokens per second has been established.

63. **What are the threshold and objective data storage requirements? What are the threshold and objective GPU requirements (Cuda, RT, Tensor cores, ram)?**
All requirements are represented by the assessment criteria

64. What is the expected deployment timeline and for how many initial units?

The timeline and units will vary per proposal. The assessment criteria notes delivery and deployment timeline as a decision factor.

65. Is the 30A already de-rated (ie 100% usable) or do we need to de-rate?

Assume you need to de-rate to ~24A continuous.

66. Will the govt consider an extension to the schedule?

No

67. Do you foresee a specific requirement around data labeling and model evaluation for this effort?

Not at this time.

68. Can you describe your expectations for operations and maintenance requirements.

The solution should not require vendor intervention to continue operating; local IT/network teams should be able to set up, run, and troubleshoot it. Limited vendor support for a short period after delivery may be requested.

69. How many GPUs are required in the cluster

Not specified to allow multiple solutions to be presented. Expect solutions to have 1-3 servers with 1 or more GPUs each (depending on model/cost per GPU).

70. What LLM types will be trained or fine-tuned (open-source, commercial, model sizes)? What maximum model parameter size must be supported?

It is not providing directions on fine-tuning use cases or model parameter sizes, and emphasized the primary focus is enabling use of popular open-source models; fine-tuning/training is secondary.

71. What batch sizes will be used for training, fine-tuning, and inference workloads? What inference latency targets must be met for 100+ concurrent users?

There were no hard targets established for 100+ user latency and welcomed vendor expertise.

72. Is there a preference for CUDA or ROCM support?

No.

73. Is there a need to connect back to a clould based environment?

The solution must be on-premises.

74. Will workloads require RAG, vector search, or embedding generation? Will users run isolated jobs, shared GPU pools, or scheduled multi-tenant workloads?

The workloads will vary to meet mission requirements. The primary use case for the GPUs is to run AI/ML models accessed programmatically by existing and new software workflows.

75. **What is the expected GPU utilization profile (continuous, burst, mixed)?**
Plan for sustained load over time.

76. **Is there an expected procurement after the initial prototype? What is the scale expected?**
Follow-on procurement scale is not yet known and depends on prototype performance; broader demand for GPUs is recognized but no specific follow-on scale was committed.

77. **Can you confirm total number of concurrent users training LLMs and the total number of concurrent inference users?**
100+

78. **Is there a preferred pricing model - CAPX vs. OPX ?**
No.

79. **Are training and inference workloads co-located or isolated on separate servers?**
Co-located, with the primary use case being inference.

80. **Does the area around the existing sites support a modular solution that can be integrated onto the current load feeding the facility?**
Solutions must be able to be installed in a standard server rack.

81. **Will users require dedicated GPU resources or shared access? What peak and sustained concurrency patterns must be supported?**
Shared access, with sustained usage.

82. **Are all workloads required to run on NVIDIA B300 GPUs? What minimum GPU count per node is required?**
No specific model or exact specifications were provided a particular GPU model or exact specs, emphasizing vendors should propose the best trade-offs to meet performance needs while minimizing cost, power, and cooling burden.

83. **Can you share a typical use case or cases?**
Hosting AI/ML models for inference use by existing and new software workflows.

84. **Do you have minimum VRAM per GPU (e.g., 40 GB vs 80 GB) and total GPU count targets, given expected model sizes? Any NVLink/NVSwitch connectivity between GPUs?**
A target of 80 GB VRAM per GPU, with a minimum of two GPUs per server and scalability up to roughly six GPUs; specific interconnect requirements were not stated as universal. Proposals are

welcome to propose varying amounts of compute and GPU to demonstrate tradeoffs (including price and timeline).

85. What GPU memory capacity is required for largest target model? Is GPU partitioning (MIG or similar) required for multi-tenant operation?
A target of 80 GB VRAM per GPU, but did not define a specific model-size requirement in parameters. GPU Partitioning not required.

86. What is the required CPU core count and type (x86 vs ARM)? Expected CPU-to-GPU ratio for preprocessing or RAG workloads?
Proposals can note CPU-to-GPU different offerings and tradeoffs.

87. What total system RAM is required per node? What minimum memory bandwidth must be met?
A target of 80 GB VRAM per GPU

88. For early stage capabilities like Crag Co LLC's deterministic GPU-coupled Phase-Change Memory technology, where technical details are shared only under NDA
Proposals are instructed to be unclassified and non-proprietary.

89. For the ROM requirement, would it be best to approach with good and best options for hardware configurations? Type and # of GPUs can effect the ROM.
Providing good/better/best hardware configuration options—especially varying GPU count (e.g., 2 vs. 6)—is a good approach for ROM pricing and scalability decisions.

90. Required network fabric type (Ethernet/RoCE, InfiniBand, or both)? Required bandwidth per node (100/200/400/800 Gb/s)?
Ethernet. OCp 3.0 is not supported.

91. Is the use case uptime-critical or are interruptions due to failures acceptable?
High availability is desirable, but mission applications are not strictly uptime-critical; planned interruptions for software updates are expected, while hardware failures are not acceptable.

92. To clarify, this solution will be drop shipped to location and all install and config will be done by the J6? No post sales services required?
The solution should be drop-shipped and installed/configured by the local J6 team, with limited vendor support expected for a few days after delivery; ongoing on-site maintenance is not expected.

93. Will there been need for GPU-to-GPU high-speed interconnects (NVLink domain size)?
If a proposed architecture requires high-speed GPU interconnects (e.g., NVLink) to meet its design, it should be included and detailed; it was not stated as a universal requirement.

94. **Will a GPU only solution that interfaces with the current compute be considered for accelerating the current workloads?**
Does not want to constrain approaches and left it to vendors to propose the right balance of simplicity, cost, and requirement fulfillment.

95. **What are the requirements for the environment (OS), AI/ML frameworks (PyTorch, TensorFlow, JAX), AMD or NVIDIA components (CUDA, NCCL, Triton, NeMo, TensorRT)?**
Software which enables GPU usage, such as driver software.

96. **Will there be a need for a cross-domain solution?**
A cross-domain solution is not required as part of this effort.

97. **How many tokens are required for the model ?**
This question is non-applicable.

98. **You mention a GPU h/w card format- is this preferred to a server based system?**
No change to the assessment criteria.

99. **What is the maximum size model that will be run on the hardware (preferable in billions of parameters)?**
Did not specify a model size in parameters, stating a preference for running the largest models feasible within trade-offs among cost, power, and cooling.

100. **What are you doing to protect against cyber-attacks? (VPN's are a placebo -- ask NSA how long it takes them to get though any SW defense.) [Systems]**
Not a focus of this effort.

101. **Will maintenance/support be conducted by the vendor? What level of support is expected? Will there be a training period + runbook dissemination?**
Ongoing on-site maintenance or support is not expected; local teams should be able to operate the solution, with limited vendor support for a short period after delivery.

102. **If we provide a software defined data storage solution on our proposal will it be utilized?**
Data storage is already provided on-site and is not a focus of this particular effort.

103. **30A limits allow for 6.6kW per rack - how much space is there to increase the number of racks in the server rooms / data center?**
Assume racks will not be added, but there may be space available on between 1 and 3 racks.

104. **What storage requirements if any are required?**

The effort is not intended to address a storage shortfall; existing storage and databases are already in place, and the need is GPU acceleration.

105. Since inference software heavily impacts GPU performance, which framework will you be using (vLLM, Ollama, Triton, etc)?

The team has used Ollama in limited contexts due to current CPU-only servers, but intends to remain flexible and may use different inference frameworks once GPU-enabled hardware is available.

106. Will the government provide a standardized benchmarking workload for fair vendor comparisons, considering software choices affect performance?

The Assessment Event is primarily a technical/architectural and value assessment rather than a standardized benchmarking bake-off, and does not currently plan to provide a standardized workload.

107. Is SOCOM interested in exploring a secure, cloud-based solution like Azure Virtual Desktop to support AI workloads as an alt to on-prem GPU infrastructure?

The solution must be on-premises.

108. How much rack space is available at the site? fitting 6 GPUs into a 3U case is very difficult for power and cooling.

Flexible - 2U, 4U, 8U cases. For this initial effort, expecting 1-3 servers with 1- 8 GPUs each (depending on model/cost per GPU). Open to multiple/tiered proposals**.**

109. Is there a current or future need for LLM processing in a tactical edge use case including a DDIL or disconnected environment?

This is not an edge AI use case; it is an on-premises solution intended to support users at one location.

110. With limited GPU-compute permutations available, what types of differentiation are you prioritizing? Tech support, deployment services, integration, or other?

Differentiators include ease of integration, performance, and balancing cost, power draw, and cooling requirements.

111. How should small businesses connect with those interested in partnerships?

The teleconference chat could be used for introductions and sharing business contact information so interested parties can coordinate outside the teleconference.

112. Due to lack of specific defined requirements can submissions contain multiple options. Ex. lower performance - low cost vs. high performing-high cost options?

Yes. Multiple options may be included, such as lower performance/lower cost versus higher performance/higher cost options.

113. Is there a ruggedized requirement for the hardware.

No ruggedization or transportability constraints are expected; the system will be installed in a server room with adequate cooling.

114. Given the requirements for enterprise products, why is the user seeking help from industry instead of buying components from GSA?

We look forward to the innovative proposals helping the government enable one of our groups with GPU resources.