

DARPA-BAA-16-52

Hierarchical Identify Verify Exploit (HIVE) Frequently Asked Questions (FAQ)

August 25, 2016

TA1

1. Q: The BAA notes scaling to 20W, is that just for the graph accelerator chip?

A: Power scaling is important for applications that TA3 identifies to be on the network edge, and in those cases the chip itself is desired to consume no more than 20W.

2. Q: Is the bandwidth goal (1TB/s) in terms of compressed data or the equivalent uncompressed data?

A: An IO bandwidth of 1TB/s is an estimate of IO performance necessary to achieve orders of magnitude acceleration of graph applications. Whether to compress (or not) is up to the performer.

3. Q: Is the HIVE chip envisioned as a pure digital design? Is non-digital processing anticipated to be in conformance with the BAA, at least for TA1? If so, will some of the primitives be evaluated for precision requirements?

A: Once the primitives are decided in phase 1, TA1 is welcome to solve/address them in the best manner possible.

4. Q: Is this seen as a standalone system or could it be attached to a conventional system?

A: The proposed hardware solution could be a co-processor attached to a conventional system.

5. Q: Is fabrication in overseas foundries a problem in HIVE?

A: No, fabrication in overseas foundries is acceptable in HIVE.

6. Q: Will a complete GDS layout specification of the chip be needed by the end of phase 2 or can that spill over to phase 3?

A: To meet the program timeline, performers should expect to deliver the GDS2 specification at the end of phase 2.

7. Q: How important is it to consider very low power solutions, for portable & in-field deployment?

A: Low power solutions (up to 20W chip power consumption) are application-dependent, though they are expected to be important for edge applications.

8. Q: Will the delivered phase 3 HIVE system be an accelerator connected/hosted by a CPU system? If so, does interconnect to the host system come into scope for TA1?

A: The I/O connection system is in scope for TA1, and co-processor designs are in scope as well.

9. Q: By which metric must the graph processor achieve a 1000x improvement?

A: The 1000x improvement is meant to be relative to power efficiency. However, the scale at which the 1000x improvement must be demonstrated is not defined. Solutions that can demonstrate 1000x performance improvement across the problem sizes of interest are preferred. However, solutions that only achieve 1000x improvement for a portion of the problem sizes of interest will still be considered.

10. Q: Are the HW performers for TA1 creating HW architectures exclusively designed to accelerate sparse computation and data movement?

A: No, the goal of TA1 is to create HW architectures and a graph processor that accelerate DoD/US government mission needs in graph computation. Graph computation (both streaming and static) is characterized by sparse computation and heavy data movement of a random nature, but can include some stages in the algorithm that involve significant dense computation and data movement. Thus, HW solutions will likely need to efficiently handle dense workloads as well as sparse workloads. The government furnished challenge problems are designed to characterize the computation and data movement characteristics of DoD and US government graph challenges.

11. Q: Are we allowed to make changes to DRAM design? If so, do they need to be compatible with DDR standards?

A: Yes, it is acceptable to make changes to DRAM design as part of a proposed solution. No, compatibility with DDR standards is not required.

12. Q: Can you clarify the 1TB/s BW per processor to memory requirement, as this seems to be on commercial roadmaps today, what size memory would be connected here?

A: This will be determined based on evaluation of the applications in phase 1.

13. Q: The BAA specifically calls out micro-code as a development expectation for TA1. The term "micro-code" has a very specific meaning in the industry. Is there an expectation that micro-code models will be exposed to other parties, or that micro-code needs to be used at all? What is the specific set of expectations around micro-code, particularly for those platforms that do not use micro-code?

A: "Microcode" was used as a generic term in this BAA. Anticipated information needed to be released to other performers is ISA, performance data, and details of the ISA sufficient to model behavior and performance.

14. Q: While the BAA calls out desired peak bandwidths to memory and between nodes, it does not discuss per-node peak computational capability in FLOPS or IOPS, nor memory capacity. What are the target goals for these?

A: Success is evaluated based on overall application acceleration, not component performance

15. Q: Must all HW solutions proposed for TA1 meet all HW interconnect, memory BW, and processing performance goals as defined in the BAA?

A: As stated in Part II, Section I.B., paragraph 4 of the BAA, "if proposed objectives are less aggressive than those defined herein, they will be considered only with justification and a sound path to meeting the final objectives, and if an appropriate risk mitigation plan is offered." HW performance goals (interconnect, memory BW, processing rate) are based on anticipated performance needed to meet mission/application acceleration and performance goals for the architecture (1000x). Following the BAA, if a proposed solution does not meet anticipated performance specifications for interconnect, memory BW, and or processing rate, then supporting evidence should be included to demonstrate that the proposed architecture can still meet the 1000x performance improvement relative to existing CPU/GPU solutions.

16. Q: Are "exotic" memories in scope (e.g. memristors, etc.)? Are neuromorphic designs in scope?

A: From Part II., Section C., paragraph 3, "the use of advanced memory technology may be necessary." Any technology that enables the hardware to meet the ultimate performance goals is within scope.

17. Q: A deliverable is listed as the architecture design, provided as GDS2 / GDSII files. Who are the recipients of the GDS2 files? What IP rights are expected as part of this? What security will be provided for delivery and storage of these files?

A: The GDS2 files will be delivered by the performer to their selected foundry for fabrication, and also to the government as a program deliverable. The IP rights and related security mechanisms between performer and foundry are the responsibility of the performer to negotiate. The government expects unlimited rights over technical data produced while performing under this contract; however, specific data rights are negotiable during the contracting phase of proposals selected for award. It is up to each proposer to clearly list all data rights assertions using the format prescribed in Part II., Section VIII.A.1. of the BAA.

TA2

1. Q: Will the HIVE graph analytics processor require a front-end with a particular operating system and programming environment w/ debugging and performance analysis tools for graph applications development?

A: The HIVE anticipates using standard environments such as Python and C along with associated graph libraries. No restrictions on operating system are anticipated in this program.

TA3

1. Q: Does TA3 have to address all 5 problem classes?

A: It is anticipated that a TA3 proposal will identify DoD usage scenarios from across the five areas with both static and streaming cases for each resulting in 10 application areas.

2. Q: Will TA3 teams be able to do research into making synthetic datasets more realistic?

A: The HIVE program is focused on realistic workloads from TA3 applications and not synthetic benchmarks. Any synthetic data must accurately reflect real workloads. It is the role of TA3 performers to determine how to develop unclassified surrogate data sets that preserve useful properties of the applications.

3. Q: What is the expected role/level of resiliency expected, especially for streaming cases where data is persistent?

A: The level of resilience will be determined by the application areas identified by TA3 performers.

4. Q: Are cyber scenarios important to the HIVE program or should TA3 look at different scenarios?

A: Cyber scenarios are viable application areas in HIVE, for example the Anomaly Detection use case given in the Proposer's Day briefing is an example from the cyber domain.

5. Q: Are proposers expected to identify static and dynamic DoD usage scenarios for each of the five areas? Will multiple problems in each area be looked on favorably? If a proposer does not identify a DoD scenario in one or more areas, will the proposal be unselectable?

A: An outstanding TA3 proposal will identify DoD usage scenarios from across the five areas and both static and streaming cases for each. Multiple possibilities from the same combination of area and static or streaming data is favorable if they emphasize different load/performance characteristics of the TA1 and TA2 outputs. One evaluation factor is the proposer's capabilities and/or related experience, which for TA3 would include insight into DoD problems to generate graph datasets. If proposers cannot identify instances of certain challenge areas from their DoD experience it may negatively affect the strength of the proposal.

6. Q: Is the information of what graph analytics the DoD is currently using in the open literature? How does one obtain such information?

A: No, information relating to graph analytics in use by the DoD is not in the open literature. From Part II., Section I.C. of the BAA under the TA3 Core Technologies subsection, "performers are expected to have knowledge of DoD systems based on prior experience..." This prior experience would provide information on graph analytics techniques and how they relate to DoD problems. TA3 will provide unclassified proxies of those algorithms and data to TA1/TA2 performers.

7. Q: Is TA3 meant for FFRDC or can a contractor bid?

A: All TA's are open to all interested proposers.

INTER-TA

1. Q: GPUs today can already provide a compute limited 16 node system with BW's that will soon approach 1TB/s, can you clarify what 1000x performance means for this system?

A: Performance improvements are with respect to the applications identified by TA3 over current CPU/GPU solutions. HIVE seeks to demonstrate approximately a 1000x improvement in power efficiency over current GPU/CPU solutions.

2. Q: How does a 100x improvement in evaluation framework relate to the 1000x goals in the BAA?

A: The 100x improvement is based on the direct (non-optimized) use of the TA2 tools on TA1 hardware, while the 1000x improvement goal should be attainable by optimized use of the tools by expert software developers.

3. Q: Please clarify importance of HIVE systems also having to perform well on traditional (dense) applications compared to conventional architectures (GPUs, etc.).

A: HIVE systems will be evaluated based on their performance across the five applications areas listed in the BAA, which will have a mixture of sparse and dense applications.

4. Q: What will be included with the government furnished challenge questions?

A: Two challenge problems will be provided to guide TA1 and TA2 development efforts and judge performance. One problem will represent a static graph problem, the other a streaming or dynamically updating graph problem. Each problem will consist of:

- a. A problem statement
- b. A data set (or tool that generates the data set) scalable from toy problems (millions of edges) to data center scale problem (trillions of edges)
- c. The computation that must be performed to solve the problem. The complexity of the computation will be scalable in some manner (ex: resolution or accuracy)
- d. The desired solution

5. Q: The BAA mentions baselining performance on a GPU. Is there a particular GPU that should be used?

A: No, however it is envisioned that proposers will compare to recent hardware so as to provide a meaningful baseline of comparison.

6. Q: What is the range of Graph sizes that are of interest? A 16 node system is still relatively small, to what size do you anticipate a fully scaled system of interest to be?

A: The goal of the 16 node system is to provide evidence that the system could be scaled in size to handle problems of at least 1 trillion edges (to as high as 100 trillion edges). The overall range of problems scales of interest are 1 million edges up to a trillion edges (with a stretch goal of a 100 trillion edges).

7. Q: BAA states that size, weight and power constraints are critical for TA1 performers to understand but it appears that this information is not provided until the end of phase 2 by TA3 performers (which is too late). Will initial guidance be given?

A: Size, weight and power constraints (SWAP) are important for a subset of applications and not all applications will be driven by SWAP constraints. Though the constraints are not finalized until the end of phase 2, we expect ongoing communication to occur during phase 2.

8. Q: The BAA briefly mentions Causal Modeling, but previously there seemed to be much more interest in dense probabilistic graph models (Bayesian Nets, Belief Networks, Ontological reasoning etc.), is that still the case? (And if so, is this only for TA3 performers?)

A: This is only one of the five general graph analytic use cases. HW and SW solutions should be broadly applicable to all five graph analytic use cases.

9. Q: In the TA3 phase 3, it calls for 16 nodes all interconnected. I am confused by the nomenclature here.
- a. Is the intention to be 16 systems which can interoperate with each other?
 - b. Or is it to have 16 of the final chips interconnected into a single graph processor to allow a larger data set to be evaluated?
 - c. Or is it simply to have 16 of these devices which can be used on a single training session with 16 operators?

A: TA3 is responsible for testing and evaluating application performance of the 16 node system in phase 3. TA1 is responsible for fabricating it. The intention is to have 16 nodes (each containing a graph processor) interconnected working on a single graph problem (generally of larger size than a single node could handle). This is similar to a cluster used in HPC computing.

10. Q: Will the graph primitives be given to us later? Can you comment on what primitives are important?

A: As specified in Part II, Section I.B.1., paragraph 5, and in Part II Section 1.B.2., paragraph 3 of the BAA, TA1 and TA2 performers are expected to create a list of graph primitives during phase 1 of the program. Additionally, as stated in Part II, Section 1.B.1., paragraph 6, “the government will provide a set of initial primitives,” with anticipated delivery of this GFE during execution of phase 1 of the program, as specified in the “Milestones and Deliverables” sections of Part II, Section 1.C. of the BAA. Applications provided by TA3 and GFE will determine primitive importance during the program.

11. Q: TA1 does not include new architecture basic tools – compilers, debuggers, etc. These are listed as TA2. How will any TA2 provide a fully optimizing tool set for every TA1 in reasonable effort? Or should these basic software tools be included in TA1?

A: TA2 is responsible for the software tools in this effort and will need to interact with TA1 performers to fully use the provided hardware from an optimization standpoint.

12. Q: Although co-design is mentioned, there does not seem to be any TA2 focus on new algorithms which take advantage of new HW. Is new algorithm R&D only for TA3 performers?

A: TA2 is focused on tool development not algorithm development, TA3 is focused on algorithm development.

13. Q: Are all TA1 & TA2 teams supposed to converge on the same set of graph primitives for the hardware abstraction layer, or can different approaches emerge?

A: Yes, it is expected that a complete set of graph primitives is identified across performers at the end of phase 1. Which primitives are implemented and how these primitives are implemented is up to the performers.

14. Q: With multiple awards how should TA2 performers approach developing toolkits compatible with possibly many TA1 hardware solutions?

A: Proposals should clearly state the number of different approaches that could be supported based on the level of effort proposed.

15. Q: Is the use of "GTEPS" used generically or as used in the Graph500?

A: GTEPS is a standard acronym contained in literature available to the general public and stands for Billions of Traversed Edges Per Second. It is consistently used in that manner throughout the HIVE program documentation.

PROCEDURAL

1. Q: Can you please clarify what is meant by "Proposer's reference number"? Page 28 of the BAA.

A: The "Proposer's reference number" is normally used as an internal control number by larger companies/organizations that send/receive large amounts of correspondence.

2. Q: What is meant by the term "lead organization"?

A: The "lead organization" refers to the "Prime Organization" that will be responsible for submitting the entire proposal. This is different from the sub-contractor, who may do work or provide support but are not the lead organization on the effort. The prime contractor will typically manage the subcontractor(s).

3. Q: What is meant by "Other team members"?

A: This refers to other organizations (subcontractors, consultants, etc.) that will contribute to the proposed effort/work.

4. Q: Will the Proposer Charts be provided?

A: Yes, they are posted on the DARPA Opportunities website under the DARPA-BAA-16-52 section: <http://www.darpa.mil/work-with-us/opportunities?PP=1>.

5. Q: Is there an Excel template for the cost buildup?

A: No, there is not.

6. Q: If TA2 performers must open source the valuable and strategic SW technology, will TA1 performers be required to open source the HW design completely? Do TA1 proposals gain advantage by open sourcing the HW in their plan? If HW is not to be open sourced, why?

A: An open source HW design is desirable but not required in this program; data rights will be negotiated on a per-performer basis. Neither TA1 nor TA2 performers are required to open source their design.

7. Q: Will you be posting responses to questions before the final deadline in case we realize we need to rephrase them if they are not understood as we intended?

A: Questions will be answered as received until the October 5th 2016 FAQ submission deadline listed in the BAA. Responses can be expected within 24-48 hours after question submission.

8. Q: Can we submit to TA3 if not all PIs have clearance? As an example, can university participants with no clearance be included as subcontractors for a TA3 prime?

A: It is permissible for a TA3 prime to work with subcontractors who do not hold security clearances; however, TA3 primes will need to be able to handle classified information and are still required to follow all applicable rules and regulations pertaining to the access and control of classified information.

9. Q: Could you clarify/explain the expectations for use of classified data in TA3?

A: TA3 requests DoD-relevant applications be identified from across the five areas listed in the BAA, to ensure realistic applications are developed, real DoD system data will be used.

10. Q: Is an abstract required to participate in HIVE? What is the expected government response time to abstract submissions?

A: As stated in the BAA, abstract submission is not required and government response on the abstract will be provided 15 days after submittal date.

11. Q: For TA1: can the abstract/proposal be for phase 1 only, or does it need to cover all 3 phases?

A: The government is seeking complete solutions across all 3 phases of the program in any of the TAs.

12. Q: When are deliverables/milestones expected?

A: Major deliverables listed in the BAA are expected to be delivered by the end of the phase in which they are described. Please see the BAA for timing information on reporting etc.

13. Q: Who pays foundry costs?

A: Foundry costs need to be included in the proposal. Ultimately, DARPA pays the costs by funding the performer during phase 3, but the costs must have been priced out and included in the proposal.

14. Q: Please comment on cross-contamination/IP issues, especially for dual TA proposers, where company A is working TA1 and TA2 and Company B is working TA1, how does Company B protect their IP?

A: The Associated Contractor Agreement (ACA) terms and conditions, to include those pertaining to IP, are negotiated between the parties (associate contractors) upon receiving selection letters from DARPA. The goal is for all ACA's to be fully negotiated/executed by the time contracts are awarded. The government will specify the parties, but not the terms of the agreement.

15. Q: Although teaming is not required, should proposers in one TA identify specific proposers in other TAs to work with?

A: No.

16. Q: Will DARPA provide some information on expected funding levels for each phase of TA1, TA2, and TA3 or other funding estimates?

A: No.

17. Q: What rights is the DoD asserting over deliverables?

A: Data rights are negotiated during contracting, please follow the guidance in Part II., Section VIII.A.1. of the BAA for more details.

18. Q: Who are some transition partners?

A: This information is available to performers.

19. Q: What is the timeline for award and contract?

A: As stated in the beginning of Part I of the BAA, the anticipated award and start date will be in April 2017.

20. Q: Are there benefits for a proposer team spanning all three TAs?

A: Which, and how many, TA's to propose to is up to each proposer.

21. Q: Given the strong focus on functional prototype delivery, what are your thoughts on inclusion of University partners?

A: We encourage the inclusion of university partners. Keep in mind the only award instruments permitted under the BAA are procurement contracts and OTA's – proposals seeking any other award instrument type, such as grants or cooperative agreements, will be deemed non-responsive to the BAA.

22. Q: Are there allowances for approved rates given to small businesses who have not had time to establish DCAA rates?

A: The use of DCMA approved rates and factors is not required in order to propose. However, for those team members who do not have approved rate and factors, all such rates and factors need to be clearly identified and clearly justified. For those team members who do not have DCMA approved rates and factors, it is recommended that DCAA Provisional Rate Agreements for the current fiscal year and two preceding fiscal years be provided, if available.

23. Q: The HIVE program seeks to create a totally new architecture for graph processing, including all software tools, in 4.5 years. Is the outcome of this program expected to be a demonstration system that is pre-production, or is it expected to be a full product offering? Are there specific expectations around commercialization for this technology, and general availability world-wide?

A: Since this is a Science and Technology (S&T) effort, the outcome is expected to be a pre-production or demonstration system. It is anticipated that the performer would continue the development of such a system for commercialization.

24. Q: Some of the deliverable expectations include physical prototypes of chips and boards. Who are the intended recipients for these prototypes? What quantity is expected? What additional terms around IP rights, firewalling, and ownership will be brought into play for these to other participants?

A: The quantity of chips and boards delivered to the Government needs to be sufficient to develop a 16 node system – so it is approach specific. The Government is ultimately the intended recipient, for purposes of the HIVE program, but it is important to keep in mind that there will be no restrictions on how the items delivered to the Government can be used or who can have access to them – they are deliverables paid for by the Government. Patent rights will be those defined in the Patent Rights clause included in each award instrument – for procurement contracts this will likely be FAR Clause 52.227-11 or DFARS Clause 252-227-7038, as applicable. Data rights will be negotiated based on what is presented in each selected proposal – proposers are required, per the BAA, to clearly identify all data or software that will be delivered to the Government with less than Unlimited Rights using the prescribed format (such restrictions will be assessed during the evaluation of each proposal, as applicable". The applicable data rights clause typically found in a procurement contract for such a program would be DFARS Clause 252.227-7013.

25. Q: For any GFE (data sets etc.) or TA3 performer input to TA1 and TA2 performers will there be any export control requirements (and confirm TA1/TA2 performers will not need any security clearance)? Will there be any US Citizenship requirements on TA1/2 performers? Will there be any government application specific requirements to TA1/2 performers (i.e. which could be construed as ITAR)?

A: Complying with Export Control is solely the responsibility of the performers and DARPA cannot offer any feedback/input regarding this issue.

26. Q: Can you provide budget guidance for individual proposals (by track and phase)? and/or overall program budget by track and anticipated number of awards in each phase?

- a. Given the as yet to be determined cross TA outputs between phases, will there be an opportunity for selected awardees to revise SOWs and reprice future phases?
- b. E.g. given the onus for TA2 performers to support all TA1 microarchitectures, this could result in considerable unanticipated work depending on the number of TA1 performers and micro-architectural details.

A: No budget guidance will be provided. Proposers are encouraged to clearly define any technical or pricing assumptions from which their proposal is based and, if agreed to by the Government, such assumptions will be made a part of the resulting award. If such assumptions, whether they be embodied in the scope of work or specifically as proposal assumptions listed elsewhere in the award instrument, prove to be invalid during performance then the parties can discuss the possibility of any necessary/resulting changes to scope, schedule or budget at that time. Such assumptions will also be reviewed during proposal evaluation.

27. Q: The schedule tightly couples TA1, TA2, and TA3 in terms of mutually co-dependent deliverable expectations. The integration is so tight that if TA2 and TA3 are held to the same schedule as TA1, their completion of a full software stack and with applications running on a 16-node platform by the end of 2021 will put heavy pressure on completing the major architecture exploration within 6 months of the program start. For example, based on typical hardware design and fabrication timelines, to make available a full 16-node validation system for software evaluation between 2020 and early 2021 would require the total completion of all architecture research at least 3 months before the end of 2017.

- a. Is there an option to extend the length of TA2 and TA3 for an additional 6-12 months through 2022, so that the software ecosystem and applications can mature on the full 16-node platform?

A: The period of performance is as specified in the BAA Part II., Section I.B.1, with a 12 month phase 1, 24 month phase 2, and 18 month phase 3 for a total performance timeline of 54 months.

- b. Is there an option to pull in the start of architecture-algorithm co-design around primitives to Q1 of 2017?

A: No.

- c. TA2 is essentially providing an optimized software layer for the hardware platform of TA1. This includes compilers, libraries, microcode and productivity tools. Given the very

tight coupling of TA2 to TA1 on a per-architecture basis, how would TA1 and TA2 work as separate selected proposals? Will each TA2 submission be expected to make that TA2 approach work on all TA1 platforms?

A: Yes, it is anticipated that TA2 performers must provide compilers, optimizers, productivity tools, etc. based on the interface provided by each TA1 performer. Communication will be necessary by TA1/TA2 performers to enable software tool development.

d. TA2 does not explicitly include algorithms and applications, which falls more closely into TA3. However to obtain full exploration of primitives and architecture opportunities, the co-design exploration of primitives needs to be done in tight communication with architects. Is it expected that proposals will include non-classified work in algorithms and applications in TA2, or is that expected to strictly be the domain of TA3?

A: TA3 will provide unclassified algorithms to TA2/TA1. GFE of two example problems will be provided at program start.

e. Due to the heavily integrated deliverable requirements and the need for significant co-design activities throughout the proposed schedule, it would seem the highest probability of success would be proposals that span all three TAs with the multiple parties sub-contracting on each TA. Is that what DARPA desires? Otherwise how will multiple TA2 and TA3 proposals integrate sufficiently deep both technically and with compatible IP rights / firewalls on this schedule?

A: Teaming strategies are left to the proposers.

28. Q: On the cover sheet for the abstract, we are supposed to specify the total amount requested from DARPA for our project. Since we will not have yet drafted the full cost proposal, knowing a precise amount will be difficult. How accurate does this number need to be? Is it o.k. if the cost value in the full proposal is different?

A: The cost estimate provided on the cover sheet of the abstract should be a rough order of magnitude (ROM) for the program cost. A full cost proposal is only required at proposal submission time, and is not expected to accompany an abstract submission.

29. Q: Is it o.k. to write a TA1 proposal only, but with some TA3 expertise on the team?

A: The composition of any proposer's team is the sole discretion of that proposer. DARPA will not provide guidance as to the make-up of a team for any TA.