Exploring Options for a Bespoke Supercomputer Targeted for Weather and Climate Workloads

Conducted for NASA by:

Bob Sorensen

Hyperion Research

October 2019

Presentation Roadmap

- Study Background and Structure
- Weather/Climate User Current and Future HPC Concerns and Insights
- HPC Vendors Inputs: Concerns and Suggestions
- Recommendations for Next Steps

Study Background and Methodology

Background and Focus of Study

NASA's earth system models have unique high performance computing (HPC) requirements, which can differ greatly from standard industry offerings

- Moreover, the gap between vendor HPC solutions and earth system models has been growing, such that these models can exploit less and less of the peak computing capability of current HPC systems

The primary focus of this study was to gather key insights, through a series of surveys with weather and climate users and potential HPC suppliers, on options available to NASA, and others, to develop a bespoke HPC system specifically targeted for weather/climate research

Study Specifics

Hyperion Research was asked to study the feasibility, including but not limited to budget, manpower, and expertise, and the potential for a shift in HPC design to better meet the requirements for NASA earth system models

- Efforts centered on the potential for the creation of a bespoke supercomputer targeted for weather and climate workloads
- Key baseline assumptions for the study included concentrating on US HPC vendors, excluding exotic hardware technologies such as quantum computing, minimizing concerns with long-term archival storage, and deemphasizing solutions based primarily on cloud architectures

Study Structure

The study was divided into two major phases targeting the weather/climate user base and a collection of potential HPC suppliers of systems to that community

- The first phase centered on a series of interviews held with expert researchers and users in the HPC-based climate and weather community to collect thoughts and insights on current and future operation requirements as well as the specific HPC hardware, software, and architectures needed to meet those workloads
  - For phase one, 15 different weather/climate organizations in the US and overseas were surveyed including ECMWF, LANL, NOAA, ORNL, UCAR, and the University of Delaware
- Many responses consisted of the collaborative efforts within a single organization and contained combined answers from multiple climate/weather experts, sometimes across organizational boundaries
  - Some survey respondents included RFPs and other technical materials outlining their current and future procurement requirements

Study Structure (cont.)

Phase two consisted of taking the results of phase one to generate a second survey of HPC suppliers and independent HPC designers to assess the challenges and opportunities of developing a bespoke HPC to meet the phase one requirements

- HPC suppliers providing input included Cray Inc, Dell EMC, HPE, and IBM
- Vendor responses consisted of either written input or conference calls and included participation by teams of in-house HPC experts, many with strong backgrounds in the climate and weather community
- Survey respondents were told that their specific answers would remain confidential, folded into larger conclusions, and reported without attribution in the final report
  - They were also guaranteed that any input provided would not carry any obligation or future commitment to NASA or any other agency

Phase One Study Highlights

Weather/Climate User Current and Future HPC Concerns and Insights

Limitations with Current and Planned HPCs

Researchers and users in the HPC-based climate and weather community had a broad list of limitations with current HPCs as well as insights on prospects for commercial HPC offerings in the next few years

- Key concerns centered on
  - Drawbacks in memory and storage latency and bandwidth
  - A lack of diversity in processor option/designs
  - The current mainstream reliance on GPUs that are not well suited to current weather/climate community workloads
  - The trend towards vendor specific interconnect options at the highest levels of computing
- Other concerns specifically mentioned:
  - Inappropriate node/CPU designs that lead to low (near 1%) efficiency on current atmosphere and ocean models
  - Overall high, and increasing, system capital and operating costs

HPC Ecosystem Concerns

Respondents offered a number of suggestions for future structural requirements needed to design and build an effective bespoke HPC centering on ensuring a well-defined, robust development effort that had applicability across a wide range of climate and weather workloads

- The effort to build such a system would need to be a community program, not a single center or lab specific effort
- The system should not be a one off “science fair project” but be a robust architectural roadmap that would persist over at least 2-3 system generations (10-12 years)
- The system would have to be competitively procured, outperform existing platforms on weather benchmarks, have lower TCO, and need to be proven reliable
- Any proposed solution would have to be configured as a true end-to-end system
  - For example, increasing the speed of the computer system by 10X without addressing the massive flow of data through the analysis and post processing system would be unacceptable

Leveraging AI into Weather/Climate Research

Another major response theme centered on bringing AI, particularity machine learning capabilities, into the programming mix, *but that opinion was not universally supported*

- Opportunities were seen for mixing machine learning (ML) with traditional computation techniques running concurrently (asynchronously) with the simulation
  - However, such usage would require a shared memory mechanism for communication between the ML and traditional computation instances
- Researchers expect to see growth in ML and increasing requirements for supporting ML software frameworks, such as TensorFlow and Keras
- GPUs might be (although are not necessarily) a significant platform for ML applications in atmospheric research
- Other areas such as distributed data processing and observations' error correction were considered good candidates for AI

Additional Insights

- Most respondents were optimistic that future performance gains were possible with straight line extensions of existing technology bases but were concerned that algorithm and related software questions could slow down the process
- Others noted that for a custom solution, potential users would need to see a substantial performance improvement (>10x) over a general purpose machine and long term support for both the architectural design and the required programming model/software stack
- One respondent stressed that some of the biggest changes always come from algorithmic changes and mapping algorithms to underlying hardware properties, so exploration of new algorithms for many components needs to be continuous
- Noteworthy cautionary comments suggested that the performance/workload data necessary to specify memory and cache characteristics for future systems is a complex and often risky exercise, and that the biggest impediment to accepting new hardware types or software is the plethora of legacy systems and training of scientists

Many Future Architectural Considerations…

Survey respondents had a long, ambitious, and somewhat diverse list of architectural considerations for any future bespoke HPC

- Heterogeneous systems should consist of mostly CPU only nodes with a smaller fraction of CPU-GPU nodes (for example, 4 GPUs per node,) otherwise the interconnect can become a bottleneck
- A key gating performance factor will be memory bandwidth per core, a measure that is at best flat, or in some cases declining, for DRAM-based CPU systems
  - The application base's dependence on memory bandwidth is such that one can predict the performance of well implemented PDE solvers on systems just based on the bandwidth
- Processors will likely have hundreds to thousands of compute cores that will require more fine-grain parallelism in the applications
  - This will drive requirements for more memory bandwidth per processor, lower latency communications, parallel I/O with high bandwidth storage and file systems
- Application scalability (which is largely on the model developers to achieve) is another key gating factor for performance improvements
  - To support scalability, the interconnect needs to be balanced relative to on node memory bandwidth and the MPI tasks must be mapped to minimize off node communications
  - Current figure of merit for a balanced system is ~12:1 (on node vs off node BW)

…and Performance Requirements

Survey respondents offered up a wide range of specific performance requirements, but two key themes emerged:

- Aggressive performance gains will be needed to meet future operational requirements
- Such gains are needed to generate interest among the potential user base for bespoke systems
- Sample Metrics
  - A minimum of 50% “computing cost neutral” speed up in combination with enhanced scalability (i.e. hardware, power, cooling, costs)
  - Multiple 10x improvement to get to global cloud resolving scales, so more interest in substantial improvements (2-10x) and not just incremental improvements (10-50%)
  - A goal of between 1,000x to 10,000x increase in computational capability within 10 years

Performance Requirements (cont.)

- Others suggested that assuming a five year time frame to get a new system and refactor code, its performance would need to be at least 32x the performance of current systems for a broad swath of the community to get wide ranging interest in porting to a new system
- Despite an emphasis on HPC performance increases from survey respondents, there was only minimal mention of targeted benchmarks to gauge performance on existing and planned key applications

Future Software Considerations

Respondents offered a wide list of software considerations as well. Two key themes were mentioned by a number of respondents:

- A wide range of existing applications need to be rewritten to take advantage of new hardware and software trends
- There needs to be a strong base of software development tools to support those activities
  - Many expect that this process will be complex and met with some resistance
- Key comments:
  - In general, existing applications are poorly architected and largely designed for traditional, low core count CPU processors
  - Need reliable performance analysis and debugging toolsets, assuming that existing code can be ported

Phase Two Study Highlights

HPC Vendors Inputs:

Concerns and Suggestions

High Interest in Exploring a Flexible Bespoke HPC Design…Within Limits

All of the HPC supplier survey respondents indicated that they would be interested in exploring the development of some form of a bespoke HPC for NASA

- However, the level of commitment and the degree to which a system would use special purpose or customized hardware and software varied greatly
- Most indicated that the option for building a special purpose one-off system or even a series of such a system over time could not be economically justified regardless of the amount of NRE that accompanied such an effort unless that particular design had value within the wider weather community at a minimum or across a number of complementary verticals
- Some respondents indicated that no amount of NRE would justify the development of a one-off system as the opportunity costs would always be too high

But Vendor Optimism was Widespread

- Most respondents noted that with increasing user demand for systems that can support a suite of diverse workloads, such as big data and machine learning jobs along with traditional modeling and simulation, the options to bring together the exact hardware and software configuration to meet specific workloads have never been better
- Most respondents indicated that they would be willing to work with NASA to develop a framework that can meet the bulk of NASA requirements
  - But only on the condition that they would also have general applicability to a wide base of potential users
- Respondents indicated that such efforts to blend NASA and more general HPC market requirements would in the long run benefit NASA procurements as they would be much more in step with more varied, timelier, and ultimately less costly HPC mainstream
- One respondent noted that any efforts to develop system software that provided data services to manage multiple tiers of storage with different access and performance characteristics would be useful in a wide variety of market segments

HPC Design Flexibility Will be Required

A key point made repeatedly by respondents was that flexibility in HPC design through either a reliance on open systems components or carefully chosen proprietary designs could significantly ease the burden for vendors attempting to craft a system that could meet NASA requirements

- Hardware flexibility could be best achieved through the right choice of memory fabric so that any processing element can be used and scaled as long as it can speak to the fabric
  - Gen-Z was cited as a promising open system bus that can support a wide range of memory access modalities
- Processor selection was viewed as a topic open for discussion
  - Intel and AMD X86, ARM, and related ARM vector processors such as those available from Fujitsu were all named as potential processor choices
  - ARM presents some attractive options for a custom design that would allow for targeted core count to memory bandwidth ratios
- Regardless of processor choice, a careful selection of processor SKU would enable the tuning of memory and interconnect bandwidth per core to maximize these when customer workloads or benchmarks show that these are performance limiters
  - The careful selection of processor SKUs to provide needed memory bandwidth would enable NASA to exploit new, rapidly emerging storage technologies including high bandwidth memory (HBM), on package memory, and nonvolatile memory as they become more available

Consider GPUs or Other Accelerators

Finally, almost all respondents stressed the need for further exploration into the use of GPUs and other accelerators into any future bespoke HPC.

- Their potential for significant performance gains in some traditional modeling and simulation jobs, as well as the promise of bringing AI techniques into some of NASA jobs, was seen as simply too great for GPUs and accelerators not to be considered

Widespread Concerns About Existing Code Base -- Modernization is Key

By far, the major concern of respondents interested in working with NASA centered on NASA's need to modernize their codes to better capture improvements in both high end hardware and software.

- Almost all respondents cited the critical need for code modernization and refactorization, while acknowledging that such a task will not be easy

- Any available NRE for a bespoke system should be committed to code optimization
  - In two or three years, there should be interesting architectures suitable for NASAs workloads, but there will still be a significant gap in software capability
- Ultimately, the barrier to achieving higher performance will not be due to hardware and the associated slowdown of Moore's Law, but instead due to the resistance to refactoring code
- Indeed, the big gains in performance will increasingly come from software
- Of all limiting factors in such an effort, the cost, time and talent needed to recast legacy code to use new technology can be the overarching issues
- There will only be escalating costs to support an increasingly obsolete programming paradigm, and the first step to addressing that reality is to 'bite the bullet' and embrace a comprehensive software rewrite plan

Widespread Concerns About Existing Code Base -- Modernization is Key (cont.)

- Cloud service providers were cited as leading the way in driving continual code refactoring as they provide new software releases on a regular basis, allowing for the quick harnessing of any new technology gains, and NASA could learn much from their processes
- Some respondents noted that code modernization efforts would reduce the need to hire or replace the declining base of uber legacy code experts such as those capable in Fortran, with a range of programmers with skills in newer programming paradigms
  - Warnings were issued however, that hiring such experts can be expensive as they currently are in great demand in both the HPC and general IT community

Gleaning NASA Requirements Are Problematic

Respondents indicated that determining exactly what such a system would look like could be challenging:

- NASA has a wide and varied range of workloads that would need to be accurately characterized and combined into a comprehensive set of hardware and software requirements
- Some respondents expressed concern that there would be 'too many cooks in the kitchen' to develop a well-defined set of vendor requirements for this community
- For any system under consideration, vendors indicated they would be chasing a moving target, and vendors would want a better understanding of what the end goals were and how best they can meet those requirements

Improved Benchmarking Efforts Are Vital

On performance metrics, respondents indicated that benchmarks like LINPACK and HPCG were not sufficient to guide them in an overall system design.

- Many indicated that benchmarks using mini application suites, testcases, or even full applications would be the best way to determine the various strengths and opportunities of any HPC design under consideration
- Questions to be addressed included: what are your current problems, what are you trying to accomplish, what steps have you taken to profile your ecosystem for bottlenecks, etc.?
- Respondents stressed the need for better analysis of existing NASA workflows and key applications to trace the use and movement of data to design systems that maximized the value of data while minimizing its movement

Upfront Codesign Efforts are Essential

All respondents strongly indicated that the first and perhaps most important step in the development of a bespoke HPC should be efforts centered on pre-RFI vendor and NASA codesign or related team activities

- Host a design conference open to all vendors that fosters frank dialogues about existing and planned codes, hardware and software requirements for those codes, and any other special considerations unique to the NASA workloads
- Any codesign activities should not target current technology or products, but instead consider what will be available three years hence
- Efforts should be made to engage the wider weather/climate community in any such codesign activities as well as to extend outreach to those outside the sector who are also looking to acquire systems that are not 'straight out of the catalog"

Upfront Codesign Efforts are Essential (cont.)

- Engage vendors to allow them to assist in analyzing workflows and applications to help them better refine their system capabilities to meet those needs
- Offer expert vendor staff to engage with NASA application developers and computational scientists to develop representation test cases and appropriate benchmarks to better understand the constraints and opportunities for performance improvements
- Work with relevant NASA experts to collect necessary information about existing and planned application workloads to better characterize the potential of different anticipated technology developments
- Use new highly productive programming languages that have been developed outside of the traditional HPC communities and new distributed computing methods that are revolutionizing many tasks
  - However, reworking workflows to use these new hardware and software technologies will require substantial effort in both programming and validation

Recommendations for Next Steps

Recommendations

Based on the information and insights gathered in both phase one and phase two of this study, there are a number of recommendations for next steps for a NASA initiative to plan and procure a series of HPCs that effectively meet NASA computing requirements in a cost effective manner.

- Such efforts can be helped considerably by enlisting the insights of one or more major HPC vendors, all of which showed an interest in working with NASA on such a project

Improve Benchmarks

NASA has a wide and varied range of workloads that would need to be accurately characterized and combined into a comprehensive set of hardware and software requirements.

- NASA planners need to more accurately assess the range of existing and planned workloads to better provide specific hardware and software requirements
- Key considerations involve an analysis of
  - Processor core counts and memory bandwidth
  - Special memory frameworks such as HBM and persistent memory
  - GPUs and other accelerator use

Improve Benchmarks (cont.)

The results of a comprehensive requirements analysis can then be used to compose benchmarks that consist of mini application suites, testcases, or even full applications to help determine the various strengths and opportunities of any HPC design under consideration.

- Such benchmarks would also be highly beneficial in providing potential vendors with insights on high-priority features and performance expectations

Embrace New Workloads

NASA should move to extend their base of HPC workloads beyond traditional modeling and simulation to increase performance on existing applications and to expand capabilities in new ones. They include:

- AI based programming not only in machine learning application, but also to accelerate traditional modern and simulation jobs
- Big data jobs, both batch and near real time that could be a mix of large storage volumes along with high bandwidth data inputs from large arrays of edge connected sensors
- Hybrid on premise/cloud computing platforms to help absorb large fluctuations in workload, to test out and perhaps take advantage of new hardware opportunities quickly, and to tap into the growing base of CSP resident HPC software

Conduct Codesign Conferences

NASA could organize a pre-competitive codesign conference or continuing series of conferences that bring together interested commercial vendors to discuss options and opportunities based on the analysis described above.

- A key goal of these co-design conferences would be to build long-term relationships with key vendors to generate a strategic multi-generation procurement plan for HPCs instead of targeting a one-time procurement
- This would be an excellent opportunity to reach out to USG organizations that are also in the process of refactoring/modernizing their code as well as facing similar HPC hardware base issues
- Likewise, similar collaborations could be established with both domestic and foreign weather/climate HPC centers

Codesign Agenda Items: Explore New Options in the HPC Planning Process

- Generate relevant projections of advances in future HPC hardware and software from a wide base of HPC vendors and users that will better align with NASA requirements but that are in keeping with larger commercial trends
- Work with vendors, many that have significant expertise in profiling and tuning applications to systems using future technologies, in helping to develop a robust methodology for making performance predictions from NASA benchmarks
- Enlist vendor insights and expertise on the best ways for NASA to begin refactoring and modernizing their current software base to best take advantage of existing and planned commercial technology and HPC products

Codesign Agenda Items: Seek Ways to Leverage NASA HPC Procurements

- Explore opportunities to help leverage any NASA hardware or software developments into the larger weather/climate HPC ecosystem to help vendors justify the use of targeted hardware or software
- Likewise identify and include other HPC verticals that share some common mission elements with NASA to generate wider interest in key developments, build economies of scale, and offer additional financial incentives for commercial HPC vendor participation

Codesign Agenda Items: Energize Code Refactoring/Modernization Efforts

- Enlist vendor insights and expertise on the best ways for NASA to begin refactoring and modernizing their current software base to best take advantage of existing and planned commercial technology and HPC products
- Enlist the aid of other USG organizations that are also in the process of refactoring/modernizing their code, especially those looking to new GPU solutions
- Engage the growing cadre of cloud service providers who increasingly are on the forefront of new HPC hardware and software techniques, especially in the areas of modular software development and deployment

Finally

Respondents from both phase one and phase two of the survey were optimistic about NASA's efforts to confront their existing hardware and software limitations.

- They indicated that many of the suggestions offered stand a good chance on ensuring that NASA has access to the kind of HPCs that will be needed to continue their world class research in key climate and weather research and operations

HYPERION RESEARCH

Thanks

bsorensen@hyperionres.com