Tips / Techniques for writing an RFP for HPC

Presented at

PEARC19

By Kyle Sheumaker, President and CTO
Advanced Clustering Technologies

advanced clustering
technologies, inc.

Outline

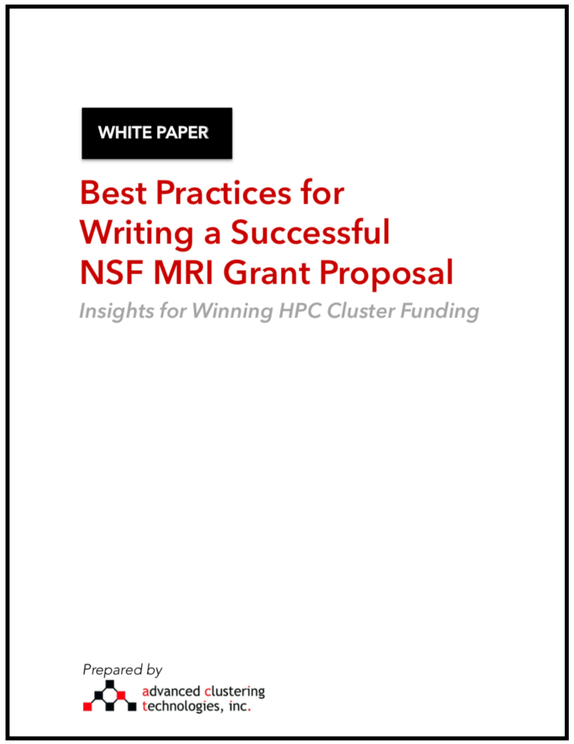
- Why we are giving this talk
- Why the details matter
- Timelines
- Technical specification guidance
  - Node count
  - CPU
  - Memory
  - GPU
  - Fabrics
  - Storage
  - Datacenter
- Q&A

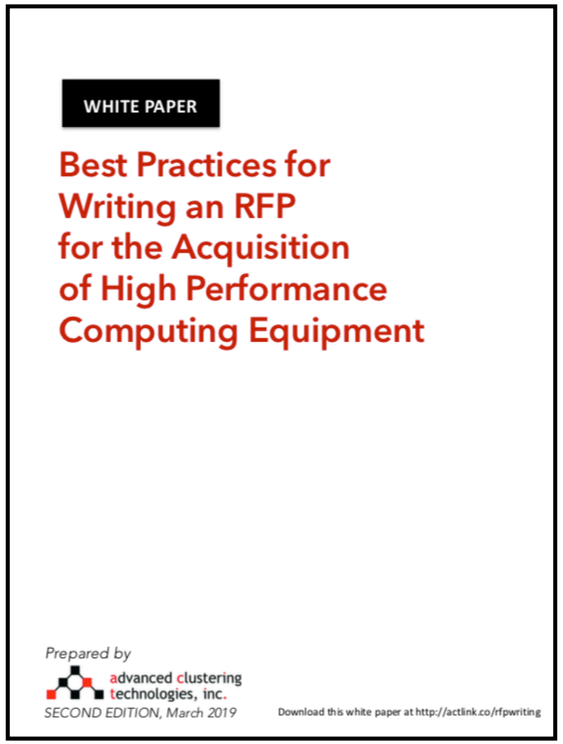
Overview

- We've been providing HPC solutions since 2001
  - Worked with hundreds of Universities from grant writing, RFPs, to installation
- The focus of this talk is on customers that are new to purchasing HPC systems
  - Buying systems in the $100,000 ~ 1 million dollar range
  - Never made a large scale purchase at their University before
  - Unfamiliar with the process

White Papers

- Created two white papers to help with the process
  - White paper on writing a grant proposal
  - White paper on writing an RFP
- Interviewed and got feedback from customers on grant writing as well as the RFP process
- Both freely available on our website





http://actlink.co/compendium

advanced clustering
technologies, inc.

Many aspects to RFPs

- Institutional rules and regulations
- Legal requirements
- Respondent qualifications
- Timelines
- Technical Specifications
- Scoring of proposals
- Acceptance testing

This talk will focus on these two topics

advanced clustering technologies, inc.

Timelines

- The RFP process can be very time-consuming
  - Preparing specifications
  - Answering Q&As
  - Evaluating responses
  - Issuing awards
- Meanwhile
  - Technology is always changing
  - Prices are fluctuating

RFP Data from 2018

Most RFPs include a schedule of the process

48 the average *estimated* number of days until award as published in the RFP

80 the average *actual* number of days until award

115 longest number of days between RFP issuance and award (they estimated 55)

8% the number of RFPs that met their published deadlines

advanced clustering technologies, inc.

RFP Q&As

Lack of detail requires lots of time spent on Q&A

38 **average number of questions asked during the Q&A period**

98 **most number of questions asked**

4 **average number of amendments issued to fix technical issues in RFP**

advanced clustering technologies, inc.

Technical Specifications

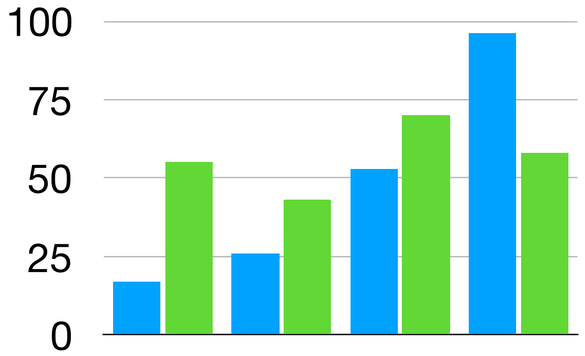
Node count

You must share something SPECIFIC:

Your Budget

Total # of cores or nodes sought

Benchmark requirements



advanced clustering technologies, inc.

Node count

Compute Nodes (The number of nodes is dependent on the proposer. The total project cost must be equal to or less than the total project cost.)

The total project cost was not included in the RFP

The new machine will have an architecture that scales well. That is, in general, for a given application and data set, the optimum number or cores will be greater than our current machine.

No list of applications or data sets were included

Node count

The number of compute nodes in the configuration should balance maximal core count and performance while staying within the total combined budget for the entire system of ~$1.3M.

A total budget for the system was specified

There are five classifications of **systems:** *head, storage, gpu, fpga, and lustre nodes*. There will be a cluster of nodes providing at least -1000 total processing cores, 20 *storage* nodes, 20 *gpu* nodes, 2 *fpga* nodes, 4 *lustre* nodes, and 1 *head* node.

Breakdown of system types and core count was provided

The aggregate performance of all servers (called *base*) must be at least 25 TFlops on the [redacted] supplied [redacted] application codes.

Benchmark performance requirement

CPU

- **If you've done testing, and analysis and know what you want:**
  - Be as specific as possible
    - Core counts
    - Frequency
    - Or SKUs, etc

- **If you are want options:**
  - Provide a metric or information of what is important to you:
    - Core count, clock speed, floating point performance, a benchmark, etc
  - Avoid using phrases like "or better" or "minimum" without quantifying what is better to you
  - If RFP scoring is mostly based on price, expect to get the minimum requested

CPU

CPU: Compute nodes should have dual socket x86_64 (Intel Skylake; provide pricing for multiple options, if available.); Dual port 10Gb SFP+. Configuration

There are over 50 dual socket Intel Skylake CPUs ranging from ~$200 - $10,000/ea

Processor: dual-socket; any Intel Xeon SP CPU that is suitable for MPI workflow; proposals with various suitable CPU SKUs are encouraged and welcome.

There were no more details on jobs being run other than "MPI workflow"

A node should be dual processor with each processor having at a minimum 10 cores.

Multiple 10 core CPUs are available with prices a range for frequencies and features, pricing from $500- $1300/ea

CPU

Dual-socket CPUs (Intel/AMD/POWER), similar to Intel Xeon Gold 6230 or better (in terms of price/performance as determined by SPECfp_rate)

Specified a target CPU, as well as open to options - including a benchmark how to compare alternatives

Intel (or equivalent) Xeon Gold 6130 (or equivalent) processors. If you would like to specify an alternate processor that provides better performance for our dollar, please provide it as a second, optional quote. Single core CPU performance is not critical for this system – overall throughput (floating-point and integer) performance per dollar is most critical.

A breakdown of what is important in CPU selection is outlined

advanced clustering
technologies, inc.

Memory

- Memory is usually the second (or sometimes the most) expensive item in the system
- Understand the memory requirements of the platforms you are requesting
  - Number of memory channels can vary (currently 6 on Intel 8 on AMD)
- Performance optimized sizes for Intel systems: 96GB, 192GB, 384GB, etc.
- For AMD systems: 128GB, 256GB, 512GB, etc.
- If allowing flexibility in CPU core counts, memory per core is usually the best approach

Memory

CPU: Dual CPU (Two(2) quantities of CPU) per server similar to Intel Xeon Gold 6130

RAM: Minimum 128GB RAM per server. Memory

128GB is not an optimal configuration for this CPU - leaving memory channels un-occupied

(2) Intel Xeon 6132

192 GB DDR4 ECC RAM - upgradable to at least 384 GB; include upgrade price to 384 GB

Requested appropriate memory, and upgrade path

Nodes must contain at least 4GB/core of DDR4 equivalent or better memory with appropriate speed.

Memory per core specified

advanced clustering technologies, inc.

GPU

- Lots of variants in GPU price
  - RTX cards ~ $200-$1000
  - Tesla ~ $9,000
- RAM on GPUs varies and can be ~ $1,000 price delta per GPU

- GPU systems are designed to support a certain number of GPUs
  - Indicating the number of GPUs per node is important
- Is future GPU expansion a consideration?

GPU

These nodes should also have NVIDIA V100 GPU processors with 32GB RAM, connected by SXM2 / NVLink.

Great detail on the type of GPU, but qty of GPUs was never mentioned

GPU: Each node should have no more than two NVIDIA GPU cards.

Specifies qty but not which cards

Fabrics

- Fabric is an important aspect to system performance
- Different configurations can have widely varying results and prices
- Knowing users performance expectations helps design an appropriate fabric
- Areas to highlight in your RFP text:
  - Speed
  - Subscription rates
  - Expansion plans
  - Physical layout

Fabric

5. EDR Infiniband (or equivalent) interconnect, including cables, as a separate budget option (2:1 oversubscription acceptable).

Specified a speed and subscription rate

2.2.4.1) A high-throughput low-latency interconnect network.

2.2.4.2) Fat-tree topology with no more than two-tiers.

Specifies number of tiers, but not subscription rate or speed

Storage

- Technical specifications for storage could be it's own talk
- You should at minimum include:
  - Filesystem
  - Reliability level (mirrors, RAIDs ,etc)
  - Usable space
- Performance expectations
  - Throughput and/or IOP/s
- Drive choices
  - Hard drives
  - SSDs (NVMe, or SATA/SAS)

Storage

Parallel Filesystem (e.g. BeeGFS, or equiv) capable of 10 GB/s and 25,000 operations/second and total usable capacity of > 500TB. For BeeGFS, two mirrored metadata servers. For other storage options, must have redundancy on metadata.

Provides detailed specifications of the storage solution they need

A parallel, high-bandwidth file system with at least 50TB, and the ability to add additional storage as needed. Lustre file system is preferred.

Specified size and filesystem but no performance expectations

Storage

In RFP

- 1.2PB Usable with option to easily expand the storage space as needed.
- Performance should scale with expansion.

In Q&A

Target performance required for storage solution?

We do not have a required target performance.

1.2PB without performance targets can produce widely different responses

Data-Center

- Provide details on where the system is going to be located
  - Photos and diagrams and floor plans are great
- Power consumption per node is increasing drastically. Important to know:
  - Voltage / Amperage available per rack
- Maximum power density per rack supported by cooling
- Location of racks (in a contiguous row, etc) will impact fabric design
  - HDR200 cable prices: 2 meter - $180, 5 meter > $1,000
- If using your equipment (racks, PDUs, etc)
  - Provide mfg, model #s, etc.

Data-Center

- One (1) vertical ServerTech PDUs (Model: CS-48VYY458A1) per rack, each with 36 x C13 receptacles and 12 x C19 receptacles, balanced evenly across three modules
- Maximum 36U of usable space per rack
- Maximum 17kW total available power per rack
- Maximum 20kW cooling capability per rack
- Prefer all/most equipment within a single rack has power supply unit(s) on the same side of rack (left or right)

Detailed space and cooling information

The HPC solution must be installed into existing racks in the Data Center at [redacted]

5.12.1.1 Liebert/Knurr Racks – Standard 42U -- Knürr Miracel®
Server Rack Model #KMK6A112000954S Serial

Model # for existing racks provided

advanced clustering technologies, inc.

Data Center

Racks available = 16 in two rows

Server racks are 19" 42u EIA-320 compliant, 1200mm "deep" racks

Cabling trays are over the racks

Individual racks cannot exceed 2500 lbs, per raised floor tile maximum weight.

Cooling capability = ~250kW; ~71 Tons

Maximum cooling capacity per rack = N/A; hot aisle containment

Great details on racks, and cooling available

- Rack integration and vendor installation

RFP doesn't give any more detail on power, cooling, rack requirements, etc.

1a. How many L6-30 power receptacles are available?

We plan to have two 20a 3-phase outlets available.

Tips

- CPUs make up the majority of the system costs.
- In general, older CPUs don’t get discounted in price when new ones are launched
  - New CPUs will be launched at the same price points with better performance
  - Best to understand the timing of launch to get most performance per $
- To get the best pricing on CPU is to include alternative options
  - When there are alternative options there is a greater chance that the CPU manufacturers will provide additional discounts
  - The more information in the RFP makes getting bigger discounts easier

Tips

- Technology is always changing, therefore It's important to utilize resources to help you stay on top of changes
  - Working with vendors before RFP issuance can help refine specifications and understand what's available within budget and and compute needs

- If you are unsure what components are correct for your requirements, there are many opportunities to test code from both Vendor or other institutional systems
- Roadmaps for upcoming technologies are under NDA, but are often available pre-launch from vendors and/or manufacturers - this can help you plan your purchase to include the latest technology

Q&A