Benchmarking and Evaluation, a vendor’s perspective

March 15th, 2016

NEC Deutschland GmbH
Intention

- I want to explain certain difficulties on the vendors’ side, no offense intended.
  - Examples are not meant to be complaints or accusations.
  - I do not have any specific project in mind.
  - Rather a few years of experience reading and analyzing RFPs.

- On my side I hope to get a better understanding of customers’ problems generating an RFP.
  - I have been working „on the other side“, too, but quite some years ago.

- In the ideal case we might create a better mutual understanding.
Benchmarking, why?

**Customer:**
- Identify the „best“ offer from a set of competing proposals, fair!
- This is prescribed by law, at least in the public sector.
- In industry it is prescribed by the necessity to proof the business case to the finance- or purchase-department.

**Vendor:**
- Understand the customer’s requirements
- Optimize the proposal accordingly
- This is necessary to have a chance

**Common interest: some kind of optimization**

**Necessary:**
- Well defined **rules**
- **Mandatory requirements**
- **Evaluation scheme** should clearly reflect the requirements
- All need to be **disclosed** – this is not always the case!

**Remark:** Vendors rely on the evaluation scheme to determine the configuration. (*1)
Rules and Mandatory Requirements

Definition: violation will result in an invalid offer

No doubt these are necessary, but ... they should not exclude valuable solutions.

- Example (*2): a certain expectation for some technical feature of the system is provided, two potential realizations prescribed as the only possibility, but there is a better solution to the problem ...

Describe the intention rather than the implementation!

Example (*3): Memory per core

- Suggestion: memory per socket, this leaves number of sockets per node open, or per node
- Case:
  - 2 sockets, 4 DIMMs per socket on the board (cheap, but why buy useless slots?)
  - 8 Gbyte DIMMs → 64 Gbyte per node
  - Requirement: 4 Gbyte per node → 8 cores per socket max.
  - Broadwell era: 10, 12 and 14 cores often provide best price / performance
  - → 16 Gbyte DIMMs? Expensive!
- Application needs?
Discrepancies

Discrepancy between mandatory requirements, benchmarking and evaluation scheme, and perhaps ... statements in the RFP.

- Example (*4): Evaluation scheme in favor of peak performance, while the RFP states application performance and importance of memory bandwidth
  - more general: text emphasizes a certain direction, the evaluation scheme does not reward it

- Example (*5): probably unintended, the way to win was to offer less kit way under budget
  - It did not work out in the end ... and I still don't know why
Not disclosing the evaluation scheme

Honesty: I would love to understand the intention
- There are cases ... agreed
- Particularly if the customers knows **extremely well** what he needs

Not disclosing evaluation scheme → gambling?

Please understand the problem:
- 5-8 different CPU SKUs (perhaps more)
- Perhaps 4 different options for the interconnect
- Say 2-3 options for the storage system
- Total: at least 40 different configurations to evaluate
- How big is the chance to accidentally select the right one?
Can the customer choose a SKU, an interconnect technology, a parallel filesystem ...?
- Not sure ... but if, please tell us!
- „optionitis“ ... a problem for everybody
- We often spend quite some time on this, time that could be used to optimize the proposal.

I am not sure how to cooperatively solve this
- For sure it would help to talk before you start your project
Setting up a benchmark

- If you want to help: Small, medium and big case, with SAME functionality!
- Please provide reference outputs and clear rules to validate results!
  - Validation: we might not know that a certain number in the output is the final accuracy of an iterative method
- Flexible Setup
  - E.g. code can only deal with power of two ... perhaps QCD
  - But well, if it is your most important code ... what could you do?
- Standard benchmarks???
  - Do they really help you?
  - SPEC: license cost! People mostly cite numbers from the internet!
- ISV benchmarks, or e.g. VASP
  - Again, there are license costs involved
  - often it takes quite some time to get a response, please help!
Optimizations?
- Example (*?): one vendor won because of optimizations, but another vendor offered a bit more of the same hardware, was this a benefit for the user?
  - Generally you want as much kit as you could possibly get
  - If the winning vendor made the very dominant production code 20%-30% faster ...
  - Or if the winning vendor could provide user trainings which resulted in faster execution of a big part of the workload?

X86-64, a self-fulfilling prophecy?
- Do you need alternatives?
- Sorry, that might appear a bit selfish now ...

Extrapolations? Difficult topic!
- Example, non-linear scaling because of cache effects
- Extrapolation requested which is impossible: e.g. HPCC random ring

And then again: evaluation scheme!

→ other issue: „magic numbers“ (later)
Evaluation of Benchmark Results

„Scaling“

- Problem: times enter the evaluation function → normalization necessary
- Often: all results are scaled by the fastest result of all vendors
  - Score = \( \frac{t_{\text{measured}}}{t_{\text{fastest}}} \)
- Or even using the slowest and the fastest result:
  - Score = \( \frac{t_{\text{measured}} - t_{\text{slowest}}}{t_{\text{fastest}} - t_{\text{slowest}}} \)

**in both cases: how can vendors optimize the proposal?**

- In the second case even the worst result can change the scene significantly!

**And there is a problem on the customer’s side**

- kind of „fake competition“
Evaluation of Benchmark Results, Recommendation

„Scaling“: relate the score to the reference result

- So you get some kind of „speedup“, it has a real meaning
- Fully transparent to the vendor
- no dirty tricks possible

Very often, and I think it is a good idea: „HPC throughput“

- Define a minimum performance requirement, or perhaps minimum speedup
- Beyond that: number of copies that can run simultaneously

• CAUTION: you need to know what is possible!
  - Increasingly difficult!
Evaluation of Benchmark Results, „δ Effect“

„δ Effect“ – intentionally lower case

- Most offers based on the same hardware → similar benchmark results
- Example: 75% score for benchmark, 25% score for technical merits
  - But variation of the benchmark timings minimal, say 10%
  - Score variation between 67.5% and 75% → technical merits become decisive!
- Honestly: no idea, is this good or bad, can one prevent it?
- In particular: what if the evaluation scheme is not disclosed?

Admittedly: this is gone if scaling is performed with slowest and fastest result
 Definition: a threshold which causes a severe change of the configuration

- example: IfiniBand-configurations
  - Two level fat tree, blocking 2:1, 36-port-switches

<table>
<thead>
<tr>
<th># of Switches</th>
<th>Max. Ports</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>72</td>
</tr>
<tr>
<td>5</td>
<td>---</td>
</tr>
<tr>
<td>6</td>
<td>96</td>
</tr>
<tr>
<td>7</td>
<td>120</td>
</tr>
<tr>
<td>8</td>
<td>144</td>
</tr>
<tr>
<td>9</td>
<td>---</td>
</tr>
<tr>
<td>10</td>
<td>168</td>
</tr>
<tr>
<td>11</td>
<td>192</td>
</tr>
<tr>
<td>12</td>
<td>216</td>
</tr>
<tr>
<td>13</td>
<td>---</td>
</tr>
<tr>
<td>14</td>
<td>240</td>
</tr>
</tbody>
</table>

The smaller the configuration and the coarser the granularity of some piece of the configuration ... the worse the effect could be.

No idea what could happen, worst case, or how to prevent
NEC brings together and integrates technology and expertise to create the ICT-enabled society of tomorrow.

We collaborate closely with partners and customers around the world, orchestrating each project to ensure all its parts are fine-tuned to local needs.

Every day, our innovative solutions for society contribute to greater safety, security, efficiency and equality, and enable people to live brighter lives.