A Total Cost Analysis for Manufacturers of In-house Computing Resources and Cloud Computing

Wolfgang Gentzsch, UberCloud, March 6, 2016

Many manufacturers are considering to move (part of) their computing into the cloud, usually starting with a total cost analysis, then performing a Proof of Concept with some of their simulation jobs, and successively all other CAE simulations. A Total Cost of Ownership (TCO) analysis therefore starts with the analysis of the company's current in-house compute environment, its costs, and its future needs, and mapping this scenario then into the cloud with a detailed cost analysis for current and future computing needs. Finally, a detailed comparison of the total cost of the static in-house computing environment versus the equivalent dynamic all-in-the-cloud solution is provided. We will illustrate this process with concrete calculations which we have done recently for a real use case for - say - company EXPL. EXPL shall have their headquarters in Europe and an R&D office in Asia. Finally, we present an offer for manufacturers who are interested to move part of their computing workload to the cloud, to perform a detailed cost/benefit analysis for the different scenarios: in-house versus cloud versus a hybrid in-house/cloud solution, for actual and future user requirements.

EXPL Current Computing Scenario

We begin with an example cost/benefit analysis for the company EXPL. The engineers in EXPL's European headquarters connect from their laptops to a centralized (pretty old) server with 24 cores. Additionally last year EXPL bought 64- cores HPC server with the same SW/HW architecture and setup and placed it in their R&D office in Asia. So far 2 simulation design engineers are working in Europe and 2 are working in Asia.

Usually their simulations run on 24 cores in Europe and on 32 cores in Asia and are typically lasting from 24h up to 60h. EXPL has been monitoring their European cluster during the last year and they've recorded an averaged usage of 75%.

Because this analysis just compares the two HPC system scenarios and their cost we are analyzing TCO only for the hardware costs and don't consider the software, assuming that the cost of software is identical on both solutions. Good news is that in the meantime all major software vendors now offer some type of pay-as-you-go license.

EXPL Next Steps

According to EXPL requirements, every CAE simulation, including software licenses, should be moved to the cloud; EXPL wants to have (part of) their workspace in the cloud. With this cloud strategy, in the meantime, EXPL is in very good company: a recent IDC's market study forecast says that HPC in the cloud is a $2.4B market today (in 2016), growing by 50% year over year, [8].

What is expected to happen is that EXPL will need extra computational power during workload peaks, thus renting the extra cores and licenses when required. However, they are still in a preliminary phase to quantify those peaks and, in case they will foresee a higher computational consumption for longer periods, they intend to rent extra cloud resources and purchase extra software application license(s).

With increasing the complexitiy of the models and the physics to produce higher quality parts EXPL is aiming at a 192-cores machine always on, with the option of bursting up to a higher amount (let's say 384 cores) whenever peak load demands are occurring.

With the acquisition of additional perpetual software licenses without GEO restrictions they are planning to scale up. In total only for CFD they are planning to have 5 users in Europe and 3 users in Asia.

Since EXPL has more CFD codes available they are considering to build up a (virtual) company-wide multi-purpose HPC platform. However, this should not only work for CFD, but also include all other engineering applications like FEM (where they have 6 users in Central Europe, 14 users in Ukraine, and 4 in Asia). In conclusion, if EXPL can demonstrate the validity of the Cloud solution for the CFD pilot, they are confident that the workload and the users involved will increase significantly in the coming years.

For the cloud EXPL needs uniformity. They consider having one Cloud access point where licenses are stored and all users would be able to work with. There is some concern about the connection latency from their Asia R&D office to a European based workspace. One solution could be to accelerate data transfer between the European workspace and the Asian hub, e.g. with NICE DCV, thus allowing Asia to experience smooth remote visualization, while at the same time the two Cloud locations are talking and updating each other as if the EXPL workspace is one. This configuration should be further discussed.

Summary of EXPL's current in-house situation:

- In-house CFD solvers
- EXPL Europe: 2 users, 24/core server (old)
- EXPL Asia: 2 users, 64 cores (from 2015)
- EXPL Europe & Asia: 6 perpetual 64-core CFD licenses for a total of 384 cores
- Simulations: run on 32 cores in Asia and 24 in Europe, typically lasting 24h - 60h.
- Average usage: on the old European server has been 75% during the last 2 years.

Next EXPL upgrade all-in-cloud:

- CFD: 5 users in Europe, 3 users in Asia, i.e. doubling the current number of engineers.
- All-in-Could: 192-core 5TB cloud environment, always on, increasing current compute environment by a factor of 3 (taking into account latest cluster technology).
- Including dynamic pay-per-use bursting up to 384 cores for peak loads (factor 6).

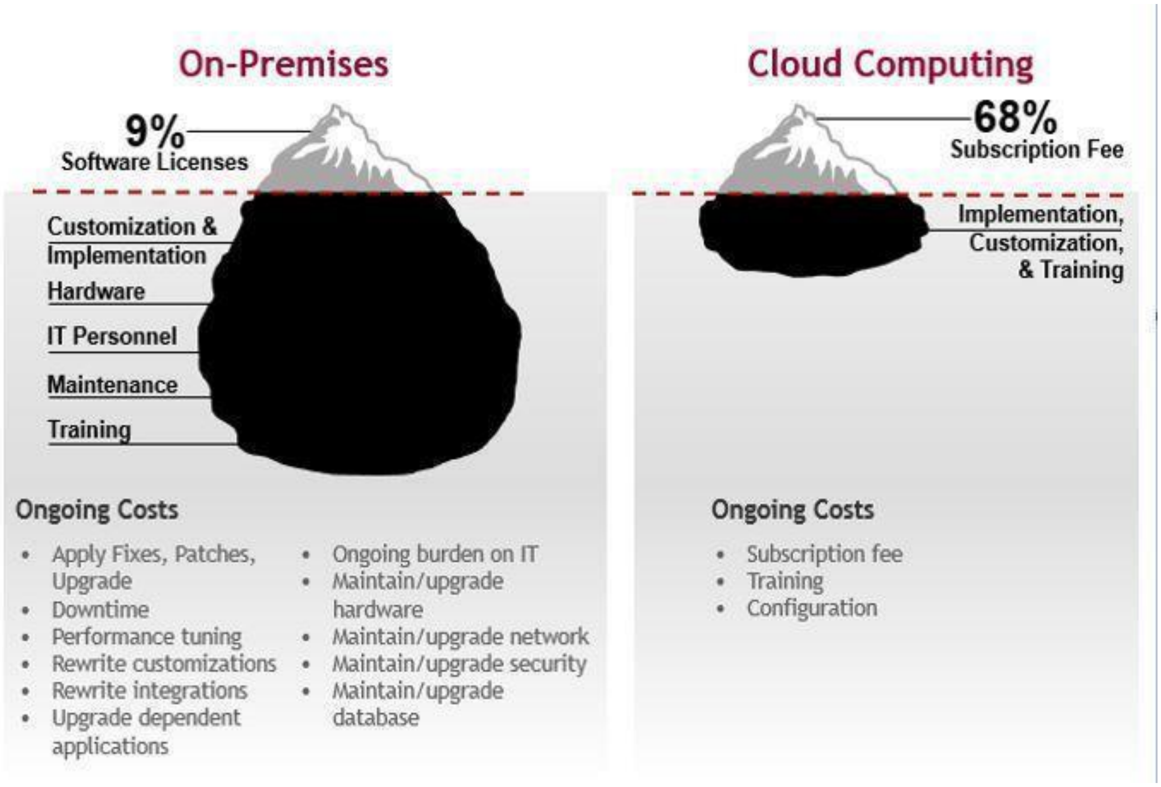
Future EXPL upgrade:

- Building a virtual company-wide multi-purpose HPC platform in the cloud.
- From single code CFD to all CFD to CAE (which will then include FEM)
- Current FEM users: 6 users in Central Europe, 14 users in Ukraine, and 4 in Asia

Total Cost of Ownership of a Compute Cluster

"One of the first things to consider is the total costs of a solution; you have to look at software, hardware, support and administrative costs. Only when armed with that information, can you make a comparison between an onsite solution and cloud-based solution costs," says Barbara Hutchings, director of Partner Relations and HPC Strategy at ANSYS, [3].

"Total cost of ownership, or TCO, is a formula that assesses direct and indirect costs and benefits related to the purchase of any IT component. The goal is a final figure that will reflect the true purchase price, all things considered," in Gigaom, [4]. The following Infographics is taken from [10].



Calculating and comparing the in-house computing environment with the cloud is a challenge because of its strong dependence on application and user requirements over the whole life cycle of the technology. Plus, some items like burdened FTE cost etc. can vary widely by company/country, and if we want to get to even more precise numbers all this has to be taken into account. But with a few assumptions we can get to the following results which are based on information from the References included below. We'll start with calculating the total cost of the EXPL cluster in case of on-premise use. First, which expenses go into the Total Cost of Ownership (TCO) of an HPC cluster? Here we use the information from an IDC study based on interview results from more than 1000 IT managers [1].

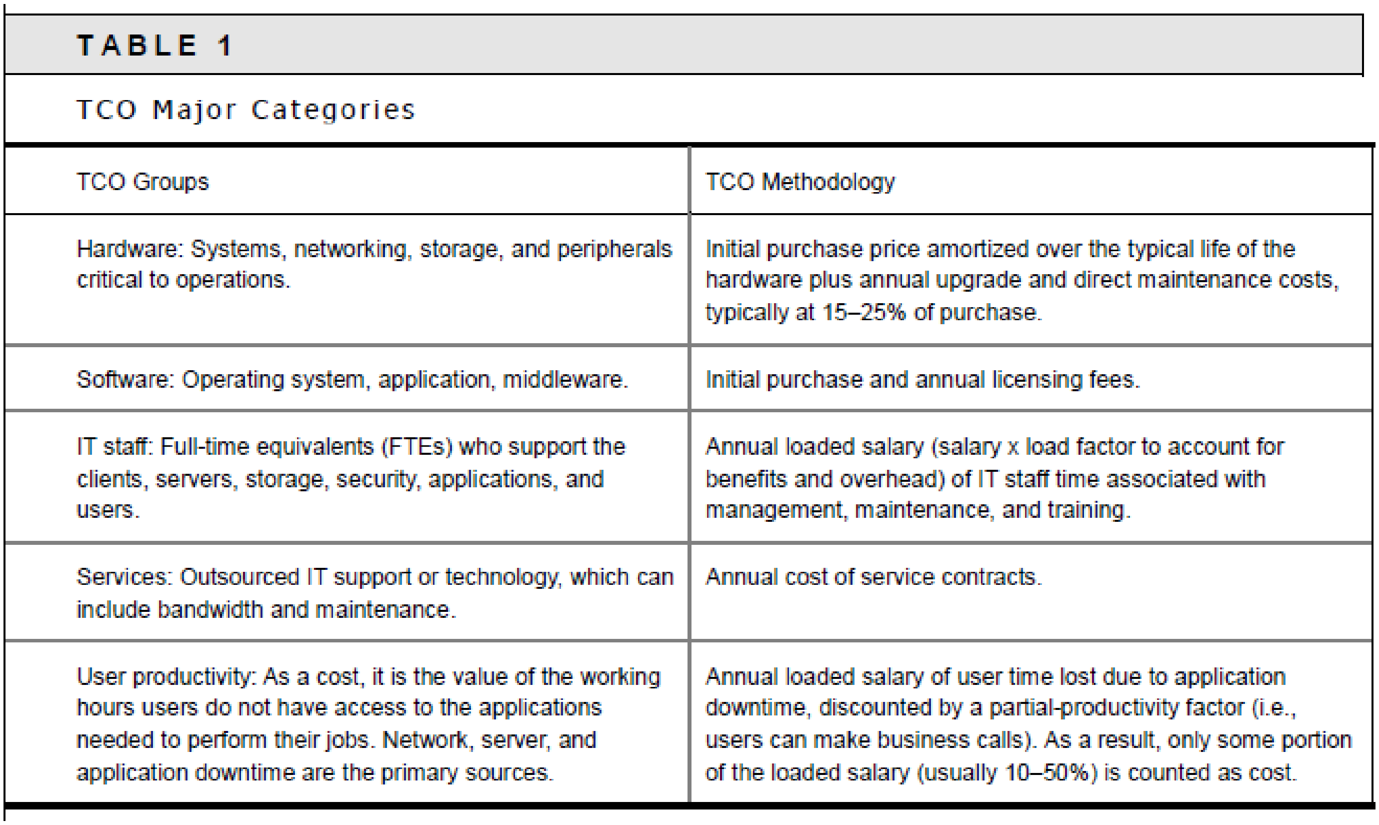


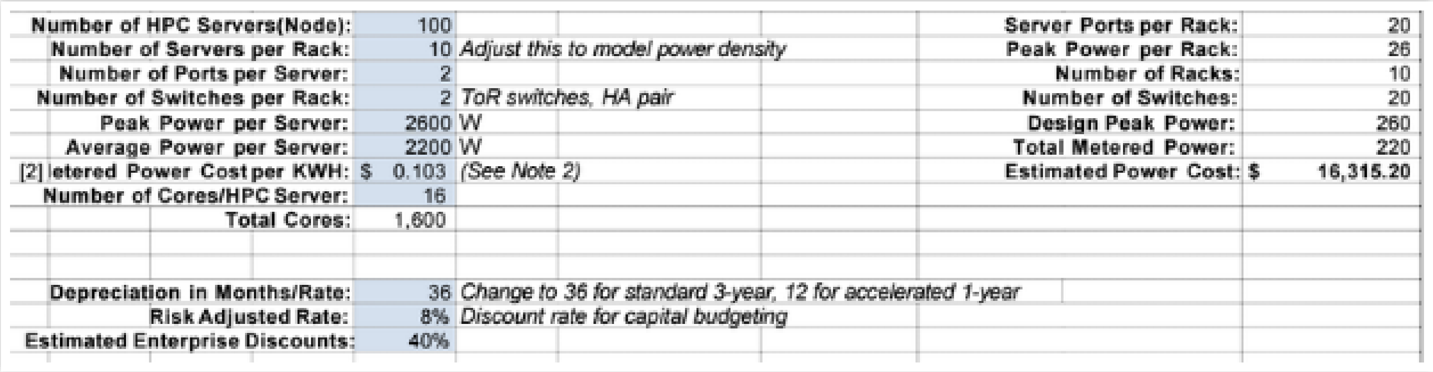
TABLE 1

TCO Major Categories

TCO Groups	TCO Methodology
Hardware: Systems, networking, storage, and peripherals critical to operations.	Initial purchase price amortized over the typical life of the hardware plus annual upgrade and direct maintenance costs, typically at 15–25% of purchase.
Software: Operating system, application, middleware.	Initial purchase and annual licensing fees.
IT staff: Full-time equivalents (FTEs) who support the clients, servers, storage, security, applications, and users.	Annual loaded salary (salary x load factor to account for benefits and overhead) of IT staff time associated with management, maintenance, and training.
Services: Outsourced IT support or technology, which can include bandwidth and maintenance.	Annual cost of service contracts.
User productivity: As a cost, it is the value of the working hours users do not have access to the applications needed to perform their jobs. Network, server, and application downtime are the primary sources.	Annual loaded salary of user time lost due to application downtime, discounted by a partial-productivity factor (i.e., users can make business calls). As a result, only some portion of the loaded salary (usually 10–50%) is counted as cost.

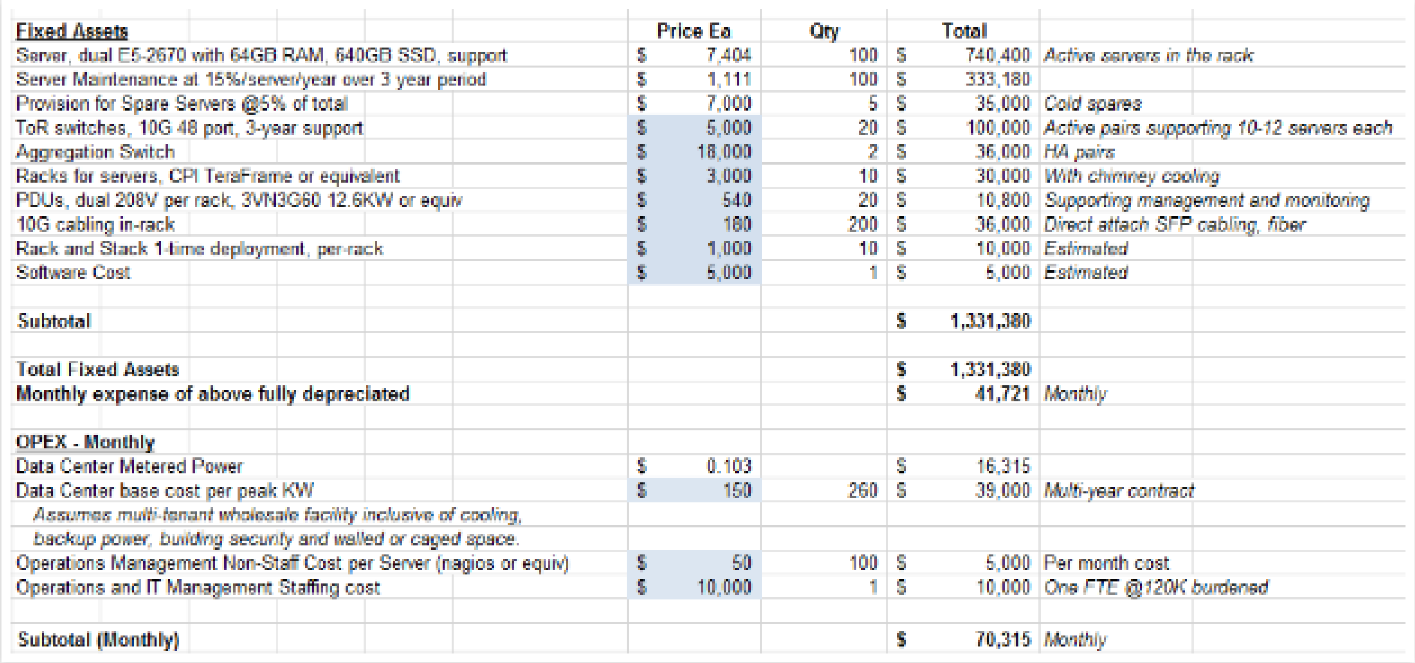
Traditionally, companies focused on the purchase price of the technology or the primary capital expense, ignoring the more significant life-cycle costs to acquire, operate, manage, and maintain that technology, and the impact on end-user productivity. More than 40 to 50 distinct factors may need to be considered when performing a detailed TCO analysis. Table 1 demonstrates how these factors can be grouped into five major categories.

To calculate the total cost of a typical in-house 100-node (1,600 cores) cluster we consider the following HPC cluster presented in [7]. Later, we will extrapolate the cost from this 1,600-core to the 192-core server for EXPL, plus an additional bursting scenario into an additional 192-core on-demand pay-per-use server, not taking into account that a smaller cluster might relatively be more expensive per core:



Number of HPC Servers(Node):	100		Server Ports per Rack:	20
Number of Servers per Rack:	10	*Adjust this to model power density*	Peak Power per Rack:	26
Number of Ports per Server:	2		Number of Racks:	10
Number of Switches per Rack:	2	*ToR switches, HA pair*	Number of Switches:	20
Peak Power per Server:	2600	W	Design Peak Power:	260
Average Power per Server:	2200	W	Total Metered Power:	220
[2] letered Power Cost per KWH:	$ 0.103	*(See Note 2)*	Estimated Power Cost:	$ 16,315.20
Number of Cores/HPC Server:	16			
Total Cores:	1,600			
Depreciation in Months/Rate:	36	*Change to 36 for standard 3-year, 12 for accelerated 1-year*		
Risk Adjusted Rate:	8%	*Discount rate for capital budgeting*		
Estimated Enterprise Discounts:	40%			

With this configuration, users are looking at close to $70,000 per month just to operate such a typical 100 node cluster in a physical datacenter, with around $16,000 of that cost going toward power and cooling. The numbers also include one full time engineer to manage the cluster in-house. Here is the breakdown of typical cluster costs (equipped with Ethernet; Infiniband will make the cluster more expensive):



Fixed Assets	Price Ea	Qty	Total	
Server, dual E5-2670 with 64GB RAM, 640GB SSD, support	$ 7,404	100	$ 740,400	*Active servers in the rack*
Server Maintenance at 15%/server/year over 3 year period	$ 1,111	100	$ 333,180	
Provision for Spare Servers @5% of total	$ 7,000	5	$ 35,000	*Cold spares*
ToR switches, 10G 48 port, 3-year support	$ 5,000	20	$ 100,000	*Active pairs supporting 10-12 servers each*
Aggregation Switch	$ 18,000	2	$ 36,000	*HA pairs*
Racks for servers, CPI TeraFrame or equivalent	$ 3,000	10	$ 30,000	*With chimney cooling*
PDUs, dual 208V per rack, 3VN3G60 12.6KW or equiv	$ 540	20	$ 10,800	*Supporting management and monitoring*
10G cabling in-rack	$ 180	200	$ 36,000	*Direct attach SFP cabling, fiber*
Rack and Stack 1-time deployment, per-rack	$ 1,000	10	$ 10,000	*Estimated*
Software Cost	$ 5,000	1	$ 5,000	*Estimated*
Subtotal			**$ 1,331,380**	
Total Fixed Assets			**$ 1,331,380**	
Monthly expense of above fully depreciated			**$ 41,721**	*Monthly*
OPEX - Monthly				
Data Center Metered Power	$ 0.103		$ 16,315	
Data Center base cost per peak KW	$ 150	260	$ 39,000	*Multi-year contract*
Assumes multi-tenant wholesale facility inclusive of cooling,				
backup power, building security and walled or caged space.				
Operations Management Non-Staff Cost per Server (nagios or equiv)	$ 50	100	$ 5,000	Per month cost
Operations and IT Management Staffing cost	$ 10,000	1	$ 10,000	*One FTE @120K burdened*
Subtotal (Monthly)			**$ 70,315**	*Monthly*

According to this calculation, the total cost for the in-house cluster, including the support staff and other utilities and services as broken down in the charts provided, is about $110,000 per month, with approximately $40,000 hardware related costs, and almost $70,000 operational (power, people, and related) costs, resulting in a TCO which is about 3 times higher than the hardware cost alone. But there is also other additional hidden cost, which an IDC study already revealed in 2007, by interviewing over 1000 IT managers, [1], see Table 1 above. This study comes to the conclusion that the TCO of an in-house cluster might even be up to 7 times higher than the hardware cost.

We now want to make a few additional assumptions to be able to calculate the cost per core hour for a 12-node 192-core cluster:

- Almost all numbers in the two tables above scale somehow with the number of servers (or cores), so we extrapolate from 1,600 cores down to 192 cores, or from 100 nodes to 12 nodes, although (as already mentined above) the cost for a smaller cluster might be relatively more expensive per core.
- With the two TCO factors of 3 and 7 from these two sources (cost tables above and IDC results from [1]) we select an average factor of 5 times the hardware cost as the TCO.

With this in mind the cluster cost for a 12-node 192-core cluster results is $41,721 / 1,600 * 192 = $5,006 for the hardware and $25,032 total cost for the 192-core cluster for one month, or $0.18 per core per hour (a metric which is widely used for renting cloud resources).
One important aspect has been neglected so far: this per-core-per-hour result is for a full (100%) utilization of this cluster! But in reality, with engineers not submitting compute jobs all the time, being on vacation, on business trips, in meeting, in weekends, and (hopefully not) sick, utilization varies drastically from almost 100% to almost 0% depending on a these factors. With project deadlines being near, cluster utilization usually goes up drastically, often with compute jobs sitting in wait queues, affecting engineers' productivity, while after project deadlines cluster utilization goes down quickly. Other observations are that with a new cluster the average utilization at first is low, with successively increasing over time, until finally the cluster is running under full utilization, finally resulting in job congestion reducing the engineers' productivity. A similar productivity loss comes also from the natural aging of the in-house cluster technology, with usually 2 – 3 technology cycles during the life time of the in-house cluster.

Getting back to our calculation, the cost per core per hour of a 12node 192core cluster heavily depends on the average cluster utilization, as follows:

Number of busy compute nodes	**1**	**2**	**3**	**4**	**6**	**8**	**9**	**12**
Resulting in cluster utilization of	8.3%	16.7%	25%	33.3%	50%	66.7%	75.0%	100%
In-house cluster cost per core/h	$2.16	$1.08	$0.72	$0.54	$0.36	$0.27	$0.24	$0.18
Cost per core/h cloud hosting*	$1.20	$0.60	$0.40	$0.30	$0.20	$0.15	$0.13	$0.10
Cost per core/h, cloud bursting*	$0.13	$0.13	$0.13	$0.12	$0.13	$0.13	$0.13	$0.13

*) Cloud pricing is market averages, including services like on-boarding, hot-line, etc. While in-house cluster cost in this example includes GB-Ethernet interconnect, the HPC cloud cost usually includes high-speed Infiniband.

Comments: The financial risk from not fully utilizing the cluster sticks always with the cluster owner, be it the in-house IT department, or the cloud provider. The cloud provider has the opportunity to rent out the un-used nodes thus driving average utilization of the cloud cluster up. And, naturally, a fully utilized cluster in the cloud has a lower total cost because usually the cloud provider gets better conditions from his hardware providers (Intel etc.) than a small or medium size company can get. And, because the cloud provider achieves (often much) better economics e.g. concerning the number of compute nodes per employee, and the infrastructure cost per node, the cost for the cluster is usually quite a bit lower in the cloud than on premise. And, finally, because the application software is used on both clusters (cloud and on-premise) in the same way, application software does not influence this TCO cluster comparison.

Additional comments on Total Cost of Ownership

IT management cost are indeed difficult to quantify: Even if ther are no full-time engineer(s) for IT, to get to a really accurate TCO one would have to record every minute from all the engineers whenever they are involved in (or affected by) IT related work, like procurement, implementation, software installation and updates, maintenance, system downtime, user support, debugging, failures and repairs, and so on. There is usually (in smaller teams) one engineer tasked with taking care of IT but all the team members are often affected as well by these activities and events, including management. And at the end it's all about loss of efficiency, effectiveness, and productivity of a company's precious engineers.

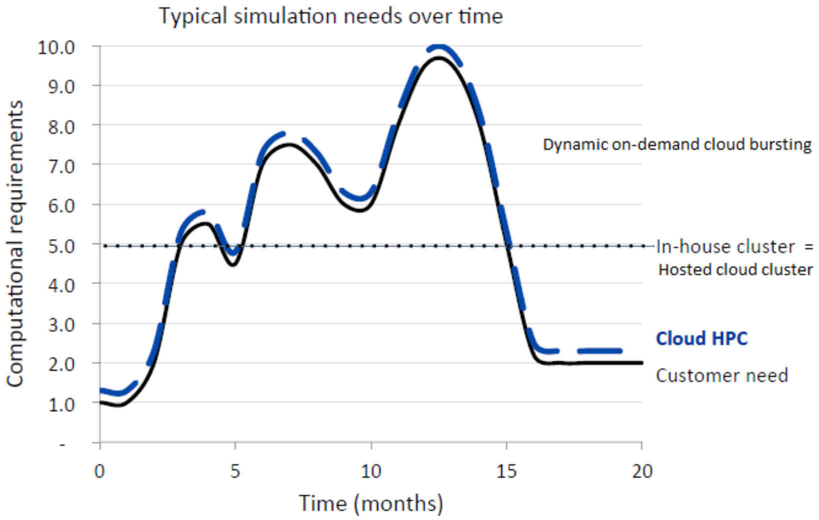
Hardware amortization in usually over three years, as suggested in IDC's TCO study. In clouds the hardware is updated at the same pace at the technology advances, i.e. 2 - 3 times a year. And, cloud providers usually keep 2 - 3 generations of hardware up and running, which results in different performance and pricings offers for the different cloud resources for the customer. And even if you continue to use your own in-house (soon outdated) hardware beyond the 3 years this results in an increasing performance loss compared to the always actual hardware available in the cloud. After three years this is easily a factor of 2 - 3, and this again has a (negative) impact on your engineers productivity. Plus now you might have to take care internally of different generations of hardware.

And what if you have **peak demands** because several engineers need to run several bigger (or many) compute jobs at the same time? And what if you could only afford to buy and maintain a mid-size system which doesn't allow for finer geometry (increasing number of cells) or more sophistic physical modeling or simply for more jobs running in parallel? This all would result in more accurate and higher quality results and a better prediction of potential failures. If you can't support this demand you would be losing competitive advantages and innovation opportunities which in fact is a huge disadvantage demonstrated by a recent IDC study about Return on Investment (RoI, [11]) from HPC. The ROI results in this study show an average of $867 revenue dollars generated for each dollar invested in HPC. And if you can't support this opportunity, you might lose part of this opportunity.

Conclusions

Considering the different cost components in the above tables and the growing resource requirements any manufacturer faces over time, one realistic and economical cloud scenario results in a combination of a hosted cloud cluster to cover the company's average utilization, plus using the cloud bursting option for additional dynamic load:

Cloud HPC provides on-demand cost-effective hardware scalability and agility for IT organizations



This all-in cloud solution, consisting of a static cloud cluster hosted in the cloud and dynamic bursting, combines all the benefits of an in-house cluster (always-on availability and reliability, application software ready to use, etc.) with the benefits of cloud bursting (scaling up and down dynamically, pay per use for above-average usage, access to always the latest technology, etc.). And for any combination of hosting/bursting the cloud solution tuens out to be always more cost effective than an in-hous cluster solution or a hybrid in-house / clolud bursting solution because, as our TCO calculations have shown above, the cloud hosting solution itself is always more cost effective than any equivalent in-house cluster solution.

This solution can be further fine-tuned to better match the overall cluster economics, by starting with a smaller hosted cluster say with 6 nodes during the time of implementation, testing, proof of concept, trials, and early production, and in a second step then (successively) ramping up the hosted cloud cluster to its full 12-node size, and beyond, in our example. When it turns out that the additional bursting becomes more intensive and thus more costly compared to the hosted cloud solution then the hosted cloud cluster can be ramped up again, and thus this hybrid schema can nicely follow the increased utilization requirements coming up in the future.

UberCloud TCO Service Offering for Manufacturers

Based on the example TCO calculations above, UberCloud has developed a TCO service for manufacturers who consider moving part of their engineering simulation workload to an HPC Cloud and who want to perform a detailed TCO cost/benefit analysis including in-house and cloud computing resources.

This UberCloud cost/benefit analysis begins with looking at the manufacturers existing resource environment and its total cost, and comparing this to an equivalent compute environment in the cloud. Next will be to look at the manufacturers future requirements (next computing demands and infrastructure, number of engineers and their tasks, applications, usage requirements in 1, 2, and 3 years, etc.).

According to these requirements we will provide a cost/benefit analysis for today and for each of the following years. The final result of this study will be a detaied cost/benefit analysis for the different scanarios, for in-house computing resources, for an equivalend cloud hosting solution, for cloud bursting on demand, and for the most cost-efficient combination of these three. Please ask UberCloud for a detailed quote, at https://www.TheUberCloud.com/help/.

References

[1] Randy Perry and Al Gillen. Demonstrating business value: Selling to your C-level executives. IDC White Paper, April 2007, https://www.platformmodernization.org/microsoft/Lists/ResearchPapers/DispForm.aspx?ID=10

[2] Steve Herbert: Comparing Costs: Dedicated HPC Cloud versus On-Demand Cluster, 2012, https://www.nimbix.net/blog/2012/07/05/comparing-costs-dedicated-hpc-cloud-versus-on-demand-cluster/

[3] Frank J. Ohlhorst: The Costs of the Cloud, Desktop Engineering 2013, http://www.deskeng.com/de/the-costs-of-the-cloud/

[4] David S. Linthicum: Cloud computing's elusive TCO (total cost of ownership), Gigaom 2014, https://gigaom.com/2014/05/09/cloud-computings-elusive-tco-total-cost-of-ownership/

[5] Timothy Prickett Morgan, Datacenter, Colo, Or Cloud – How Do You Decide? May 27, 2015, http://www.nextplatform.com/2015/05/27/datacenter-co-lo-or-cloud-how-do-you-decide/

[6] Tony Spagnuolo: The Real Cost of High Performance Computing, January, 2015, https://blog.rescale.com/the-real-cost-of-high-performance-computing/

[7] Nicole Hemsoth: The Cloud Versus HPC Cluster Cost Conundrum, June 2015, http://www.nextplatform.com/2015/06/03/the-hpc-cloud-versus-cluster-cost-conundrum/

[8] Wolfgang Gentzsch: How Cost Efficient is HPC in the Cloud? A Cost Model for In-House Versus In-Cloud High Performance Computing. February 2014, updated September 2015, http://www.theubercloud.com/cost/

[9] Chirag Dekate, Earl C. Joseph, and Steve Conway, Worldwide HPC Public Cloud Computing 2014–2017 Forecast.

[10] On premise versus cloud computing infographics: http://www.databax.co.uk/blog/the-best-cloud-computing-infographics-and-images-ever#.Vt06X5MrIUE

[11] Earl C. Joseph, Steve Conway,and Chirag Dekate, EESI-2 Special Study To Measure And Model How Investments In HPC Can Create Financial ROI And Scientific Innovation In Europe: http://www.eesi-project.eu/wp-content/uploads/2015/05/EESI2_D7.4_Final-report-on-HPC-Return-on-Investment.pdf