NIELS BOHR INSTITUTET
KØBENHAVNS UNIVERSITET

Til DeICs bestyrelse

Vedr.: Fremtidig organisering og drift af HPC i Danmark

Hermed kommentarer til de to scenarier for fremtidig organisering og drift af HPC i Danmark, som er beskrevet i de dokumenter, som fra den 18/3 har været tilgængelige på DeICs web.

Som pointeret i det notat, som vi den 12/3 sendte til DeICs bestyrelse, er en detaljeret beskrivelse og forståelse af scenarierne nødvendig for at der efterfølgende kan træffes de rigtige beslutninger, som fremadrettet sikrer optimale betingelser for brug af HPC i Danmark. De to scenarier fra DeICs dokument "DeIC_HPC_scenarier-18-03-2013.pdf",

Scenarium 1: Videreførelse af den nuværende decentrale opbygning med Tier2 anlæg
Scenarium 2: Kombination at et eller flere nationale Tier1 og et antal Tier2 anlæg

er nedenfor sat i relation til de seks faktorer, som vi tidligere fremhævede som vigtige:

1) tidsramme for installation
2) initiale omkostninger
3) dækning af spektrum af behov (type af hardware)
4) forventet performance primo 2014
5) forventede fremtidige omkostninger for drift
6) finansieringsmodel og forventede fremtidige omkostninger for investeringer, som tillader performance at stige i takt med den forventede internationale udvikling

APRIL 6, 2013

JULIANE MARIES VEJ 30
2100 KØBENHAVN Ø

TEL 35 32 59 99
DIR 35 32 59 68
MOB 21 45 49 83

aake@nbi.dk
www.nbi.dk

REF: ÅN/mb

1) Tidsramme for installation:

Scenarium 1: Et kald til ansøgninger, af samme type som under DCSC vil, efter meddelelse om bevillinger, forholdsvis hurtigt kunne føre til opgradering af eksisterende anlæg. Der vil være tale om en foregående investeringspause på mindst tre år (svarende til et tab af en faktor fire i performance forbedring relativt til gennemsnit for andre lande). De lokale centre kan derfor forventes at gennemføre fornyelser mindst lige så hurtigt, som det er sket under DCSC-tiden, dvs. i løbet af ca. 6 måneder.

Scenarium 2: Der vil her være tale om en stor investering, sandsynligvis—selv efter kompensation for pristalsudvikling—den største enkeltstående HPC investering i Danmarkshistorien. Udbud og planlægning vil og bør tage relativt lang tid. At antage, at ikke kun planlægningen og indkøbet, men også installationen, frem til det punkt hvor anlægget kan frigives til brugerne, kan gennemføres på et år ville være meget optimistisk. En mere realistisk tidsramme er halvandet år.

2) Initiale omkostninger: Hvis en sammenligning skal give mening må den gennemføres for de samme totale omkostninger. Da de samlede udgifter over 3 år i Scenarium 1 er ca. 104 Mkr, og der i DeICs materiale forklares, at Scenarium 2 (samlede udgifter til Tier-1 63 Mkr) også indebærer fortsatte investeringer i lokale anlæg, kan ensartede omkostninger nås ved at antage, at omkostningerne for lokale anlæg i Scenarium 2 er 41 Mkr, hvilket er ca. 40% af omkostningerne i Scenarium 1. De initiale omkostninger er derfor:

Scenarium 1: Som i DeICs tabel: 64 Mkr

Scenarium 2: For Tier-1, som i DeICs tabel: 40 Mkr. For lokale anlæg, ca. 40% af ovenstående, dvs. ca. 25 Mkr, i alt ca. 65 Mkr.

3) Dækning af spektrum af behov: Vi sammenligner her de dele af investeringerne, som er forskellige, og ser bort fra de investeringer i lokale anlæg, som er de samme i de to scenarier:

Scenarium 1: De lokale anlæg dækker per definition de lokale behov med henblik på Tier-2 niveau installationer. De består af en blanding af installationer til *capacity computing*; og *capability computing.*

Til kategorien *capacity computing* hører f.eks. installationer med et stort antal noder til processering af data enkeltvis, med kun ringe behov for kommunikation, eller f.eks. et mindre antal noder med stor hukommelse og et forholdsvis stort antal kerner per node. Der kan også være tale om noder forsynede med GPGPUer, til processering enkeltvis, dvs. stadig med ringe behov for kommunikation mellem noderne.

Til kategorien *capability computing* hører anlæg med et stort antal noder og meget hurtig kommunikation, men ikke specielt meget hukommelse per node. Et andet eksempel er et anlæg med kun relativt moderat kommunikationshastighed, men meget hukommelse per node. Anlæg til parallel processering med GPGPUer vil, til forskel for anlæg med GPGPUer til enkeltvis processering, have et ekstremt højt kommunikationsbehov, for at matche den accelererede beregningshastighed per node. Nogle ny-installationer vil med stor sandsynlighed vælge at bruge Intels nye MIC teknologi til accelereret computing.

Denne type af investering er *omkostnings-optimeret*, fordi man til enhver funktionalitet kun betaler for netop det påkrævede. Der investeres ikke i hukommelse, som der ikke er brug for i den del af brugerskaren, og heller ikke i hurtig kommunikation til dem, som ikke har brug for det.

Scenarium 2: Scenarium 2 vil under de givne forudsætninger, indebære et investeringsniveau i lokale anlæg, som ligger betydeligt (ca. 60%) under niveauet i Scenarium 1. Den del af Scenarium 2 vil derfor i betydeligt mindre grad end Scenarium 1 opfylde de lokale HPC behov. Hvis Scenarium 2 skal være en fordel for de danske HPC brugere kræver det så, at nytteværdien af det nationale Tier-1 center mere end opvejer reduktionen af investeringer i lokale anlæg.

Et nationalt anlæg vil ikke være i stand til at implementere det brede spektrum af omkostnings-optimeret Tier-2 funktionalitet, som svarer til den reducerede investering i lokale anlæg. Det vil ikke give mening at investere i et nationalt Tier-1 anlæg, som ikke har den hurtigste kommunikation på alle noder, og som ikke har rigeligt med hukommelse på alle noder. En påtænkt GPGPU installation på en del af noderne stiller ekstra store krav til kommunikationen, i det mindste på denne del af anlægget.

Med mindre man argumenterer for, at et nationalt anlæg vil kunne opnå betydeligt større rabatter, end de lokale anlæg har været i stand til at opnå, hvilket er svært at forestille sig, da de lokale anlæg har vist sig, at være ekstremt dygtige til at forhandle priser, så vil den manglende diversifikation af Tier-2 kapacitet altså under ingen omstændigheder kunne matches på et nationalt anlæg.

Man kan så indvende, at det slet ikke er *meningen* med et nationalt Tier-1 anlæg at opfylde de mange Tier-2 behov. Meningen med et nationalt Tier-1 anlæg er, især at tilfredsstille behovet af *capability computing*, samt at have oplæring og træning i HPC i samme center. Da vi, som forfatter dette notat, netop tilhører denne kategori af brugere, vil vi håbe på og regne med, at vores synspunkter nedenfor på disse aspekter tages særdeles alvorligt:

Et nationalt center har sin berettigelse i at dække et behov for Tier-1 kapacitet i Danmark. Dette behov skal altså være så stort, og så vigtigt, at det opvejer den uundgåelige reduktion af støtten til Tier-2 type kapacitet. Der er umiddelbart store problemer med denne forudsætning, da danske *capability* brugeres behov for kapacitet på en aldeles udmærket måde allrede varetages via vores medlemskab af PRACE og DECI, som tilbyder Tier-0 kapacitet i verdensklasse. Danske forskere har været gode til at udnytte disse muligheder, og antallet af PRACE-bevillinger, som er tildelt danske forskere stiger for hvert år.

Udover Tier-0 tilbyder PRACE/DECI *desuden*, og med meget lavere tekniske og videnskabelige krav, betydelige mængder Tier-1 kapacitet. Danmark betaler et fast årligt beløb til PRACE, som er *uafhængigt af* hvor store bevillinger, der bliver givet til

danske forskere. Bevillinger fra PRACE baseres *udelukkende* på videnskabelig kvalitet og de tekniske muligheder for at gennemføre et projekt med et 'peer review' system, med typisk 3-4 referees per ansøgning, og med mulighed for, at ansøgerne kommenterer på de indkomne referee reports.

Hvilke behov vil et dansk Tier-1 system kunne / skulle opfylde i denne situation? Der er principielt set tre muligheder, hvor størrelsen af efterspørgsel vil afhænge af access-kriterier: 1) man bruger et lignende 'peer review' system som PRACE/DECI. 2) man bruger *ikke* et tilsvarende system, men uddeler kapacitet efter mere summariske vurderinger, eller 3) man sælger kapacitet til en fast timepris, f.eks. af den størrelsesorden, som er nævnt i DeICs tabel for Scenarium 2; ca. 4000 kr/time ved 15k kerner.

I det 1. tilfælde vil nytteværdien af et dansk Tier-1 system være den, at danske forskere får *udvidet* adgang til den type kapacitet, som man allerede nu kan søge fra PRACE/DECI. Der skal så vurderes, hvor stor denne kapacitetsudvidelse ville være, og om der reelt er et behov for den.

I det 2. tilfælde kan der ske det, at når man forholdsvis nemt kan få kapacitet på et nationalt anlæg, så vil man være tilbageholdende med at bruge den sværere vej via PRACE/DECI peer review. Det vil uundgåeligt føre til en mindre kvalificeret prioritering af brugerskaren, og til at en del af kapaciteten bruges af (fra de bedst kvalificerede brugeres synspunkt "blokeres af") mindre kvalificerede brugere.

I det 3. tilfælde vil brug af systemet på et niveau relevant for en seriøs Tier-1 bruger, f.eks. 5% af de totale ressourcer, koste et beløb af størrelsesorden 1.5 Mkr/år. Ved 15k kerner svarer 5% til ca. 6 millioner kernetimer per år – en ikke ret stor kapacitet med international målstok (og heller ikke mere, end ca. 30% af kapaciteten på KUs lokale driftcenter). Alligevel vil der skulle skaffes et beløb, fra råd og fonder, som er langt fra trivielt. Det vil formentlig betyde, at de fleste potentielle brugere i stedet for vælger, at søge gratis adgang til Europæiske Tier-1 eller Tier-0 anlæg, hvor man også kan få langt større bevillinger, med langt mindre besvær og større chance for, at komme igennem med en ansøgning. Hvis man vælger denne model vil efterspørgslen på det nationale anlæg sandsynligvis blive meget lille.

Kun i det første tilfælde, hvor der også i Danmark bruges et stringent peer review system vil det danske Tier-1 anlæg potentielt kunne fungere som træning og kvalificering til PRACE Tier-0 adgang, hvilket så ville kunne ses som en potentielt vigtig fordel.

Der er, imidlertid, allerede nu adgang til topkvalificeret træning i brug af de Europæiske Tier-0 anlæg. Såvel PRACE som flere af de centre, der tilbyder et hjørne af sin kapacitet til PRACE, tilbyder internetkurser, kurser på stedet, sommerskoler, samt omfattende dokumentation, inklusive mange avancerede hjælpepakker og udviklingsmiljøer.

> *End ikke en meget stor investering i etablering af et træning og support ved et dansk Tier-1 anlæg ville, selv efter lang tid, kunne matche dette frit tilgængelige materiale fra de største HPC centre i Europa. Som medlem af PRACE har alle danske forskere og studerende gratis adgang til kurser og træningsaktiviteter.*

PRACE har også oprettet mulighed for "Preparatory Access" adgang til PRACE Tier-0 ressourcer via korte summariske ansøgninger. Sammen med adgangen følger dedikeret teknisk support. Preparatory access er netop til for at give forskere og industri mulighed for at optimere de tekniske aspekter af deres programmer, især med henblik på parallelisering, sådan at de på tilfredsstillende måde kan udnytte Tier-0 ressourcer, og gøre det muligt for flere at blive kvalificeret til at søge om egentlig og substantiel PRACE Tier-0 eller Tier-1 kapacitet.

Et nationalt dansk Tier-1 anlæg som et "trin på vejen" for potentielle PRACE Tier-0 brugere er derfor i bedste fald ikke nødvendigt, og i værste fald en hindring. PRACE giver i sig selv mulighed for at bruge langt større ressourcer. Som et eksempel har vi modtaget 180 millioner core-hours fra PRACE fordelt over de sidste tre år, hvilket svarer til et konstant forbrug på 7.000 kerner, eller i hvert fald halvdelen af den totale kapacitet i et muligt nationalt center.

4) Forventet performance primo 2014: Her skal det bemærkes, at en sammenligning egentlig burde foretages i forhold til en *fremskreven* performance for de lokale anlæg, som den ville have været, hvis DCSC-modellen var blevet ført videre, eller hvis den overgangs-

fase, som DeIC oprindeligt havde til hensigt at gennemføre, var blevet gennemført uden forsinkelse. Den investeringspause på ca. tre år, som der vil være tale om for Scenarium 1, selv hvis der bliver lavet et opslag meget snart, svarer til en gennemsnitlig performancetab, relativt andre lande, med en faktor fire, da man iflg. Moore's lov ser en gennemsnitlig fordobling af performance hver 18. måned. Vi undlader dog at gennemføre denne sammenligning, og vælger, i lyset af den forudgående diskussion af forventet tidsramme for installation, 'medio 2014' som sammenligninstidspunkt.

Scenarium 1: En realistisk og ligetil vurdering af hvilken kapacitet som ville kunne opnås i denne model kan findes ved at tage udgangspunkt i, at kapaciteten medio 2011 (den samme som nu, da der ingen bevillinger er givet de sidste to år), iflg. DeICs bilag 2 var ca. 600 Tfl, og var resultatet af løbende investeringer på ca. 20 Mkr/år. Konvertering til nypris 2011 opnås ved at gange med 2.3, da løbende investeringer af 1/3 per år vedligeholder performance ved ca. 76% af en fuld investering det sidste år, hvis der antages, at ældre udstyr tages ud af drift efter 4 år (estimatet forstyrres ikke nævneværdigt af, at nogle anlæg i bilag 1 tæller kerner med, som var ældre end 4 år i 2011). Det vil sige, at de lokale centres købekraft i 2011 var ca. 600/(20*2.3) = 13 Tfl/Mkr. Fremskrives dette til medio 2014 – altså med en faktor 4 – bliver den ca. 50 Tfl/Mkr, og en initial investering på 40 Mkr vil så svare til ca. 2 Petaflops (herefter Pfl).

Man kan ligeledes konvertere det veldokumenterede antal kerner på de lokale systemer til en pris per kerne = ca. (20*2.3 Mkr) / 25.000 kerner = ca. 1800 kr/kerne.

> *De eksisterende lokale centre opnåede altså effektivt set allerede i 2011 en pris per kerne, som var klart bedre end den, som DeIC nu bruger som estimat til anskaffelser, som vil ske i 2014.*

Scenarium 2: Iflg. SDUs vurdering af deres egen model ville kapaciteten af dette anlæg initialt kunne blive ca. 2.4 Pfl. SDUs vurdering er dog behæftet med en række fejl:

1) Der tages udgangspunkt i Oak Ridge anlægget Titan, uden at tage højde for, at Titan kom til som en *opgradering* af systemet Jaguar*,* hvor mange af komponenterne er blevet genbrugt. Et alternativt udgangspunkt ville være Blue Waters anlægget ved NCSA (Urbana, Ill.), som har kostet ca. 188 MUSD, og har en samlet CPU+GPGPU kapacitet på 4.5+12 = 16.5 Pfl. Skaleres dette til 40 Mkr bliver resultatet ca. 0,7 Pfl.

2) Selv hvis man bruger Titan som skabelon er SDUs estimat behæftet med meget store fejl: SDU bruger åbenbart 30 Pfl som performance for Titan (i det 8% sættes lige med 2.4 Pfl). Men på Titans hjemmeside er performance givet som 20 Pfl. Titan opnår denne performance frem for alt ved brug af ca. 19.000 GPGPUer; bidraget fra de almindelige CPUer er langt mindre (ca. 1.6 Pfl). Men i SDUs estimat er kun 1/3 af noderne forsynede med GPGPUer, og den forventede performance er derfor kun ca. 0,7 Pfl, konsistent med ovenstående.

> *Vælger man, i stedet for, at forsyne alle noder med GPGPUer vil der kun formelt opnås en højere performance, i det kun en brøkdel af de til en hver tid kørerende jobs reelt vil kunne bruge GPGPUer; dette ville således være en voldsom fejlinvestering.*

3) Der tages heller ikke højde for, at en meget mindre dansk investering ikke kan forventes at opnå samme prisfordele, som store nationale anlæg i USA, ej heller for at den effektive dollarkurs for IT-udstyr typisk ligger et godt stykke over den nominelle dollarkurs. Ovenstående estimat (0,7 Pfl) er derfor optimistisk, med 0,5-0,6 Pfl som mere realistiske skøn.

> *Selv om Tier-1 delen af Scenario 2 udgør ca. 60% af et budget, som forudsættes at være forøget med ca. 50-60% relativt DCSC-perioden, og alene Tier-1 delen således svarer til 90-100% af det tidligere budget, og selv om at man kunne forvente ca. en fire-dobling af performance siden de investeringer blev foretaget, som tegner sig for hoveddelen af de lokale centers performance, så er den (for fejl korrigerede) forventede Tier-1 performance således ikke væsentligt større, end den de lokale centres nuværende performance.*

Når prisen fremskrives til 2014 (en forventet reduktion med en faktor 4) er der ingen tvivl om, at de lokale centre har været betydeligt mere effektive til indkøb, end DeIC forventer, at det nationale center vil kunne være. Dette vil være tilfældet, også om DeICs prisestimat – optimistisk – fremskrives med en reduktionsfaktor svarende til halvandet år.

Tager man desuden højde for at tidsprofilen for Scenarium 2 er forsinket med ca. 1 år relativt Scenarium 1, og at kapaciteten er fordelt betydeligt mindre optimalt er konklusionen uundgåelig: Scenarium 2 kan ikke konkurrere i tilbudt kapacitet og nytteværdi medio 2014. Dette fremgår også klart af det faktum, at Tier-1 anlægget i Scenarium 2 end ikke kan overgå det *nuværende* antal kerner i de lokale centre.

De forskelle i forventet kapacitet, som vi kan konstatere, er konsistente med det faktum, at den intuitivt attraktive tanke om, at en maskine, som er dobbelt så stor bliver mindre end to gange så dyr, ikke er korrekt. Tværtimod viser det sig, at jo større maskinen bliver, jo dyrere bliver den, målt pr. FLOPS. Dette er først og fremmest på grund af omkostningerne ved skalering af netværket imellem maskinerne, og båndbredde af filsystemer, men beror også på andre faktorer, så som den større bygning, strøm-infrastruktur og køling.

5) Forventede fremtidige omkostninger for drift: Med sammenlignelig hardware vil de direkte omkostninger til strøm og køling være sammenlignelige; dog eksisterer de lokale anlæg allerede, dvs. der vil spares en del på investeringsudgifter. De lokale anlæg vil kunne forventes fortsat at kunne drives med kun en moderat udbygning af strøm og kølingskapacitet, da nyere computersystemer er mere energieffektive, sammenlignet med de nuværende installationer. I modsætning hertil vil en stor national Tier-1 installation kræve et helt nyt datacenter med tilhørende kølings- og strøm-infrastruktur.

DeIC har i sine estimater (tabeller hørende til Scenarier 1 og 2), uden nærmere motivering regnet med, at 'Datacenter-udgifter' er 10% lavere for det nationale center. Iflg. ovenstående er dette ikke realistisk; de lokale centre kan i virkeligheden forventes, at have lavere udgifter.

DeIC har også opgjort omkostningerne for systemadministration of support til at være betydeligt større i Scenarium 1 (12 Mkr) i forhold til Scenarium 2 (5 Mkr). Sammenligningen bør dog foretages for konstant samlet bevilling. For Scenarium 1 er denne som angivet i tabellen, 64+49,3 = 103,3 Mkr. I Scenarium 2 er de samlede udgifter til Tier-1 anlægget 40+23 = 63 Mkr, hvilket som nævnt ovenfor skal kompletteres med 40 Mkr i fortsatte udgifter til lokale anlæg, for at opnå et korrekt sammenligningsgrundlag. Da bemandingen på de lokale centre er meget lidt afhængig af størrelsen af hardwareinvesteringer medregnes samme udgifter til lokale centre, som i Scenarium 1. De samlede personaleomkostninger i Scenarium 2 er så 5 + 12 = 17 Mkr.

> *Selv om man vælger, at nedlægge nogle af de lokale centre, skal der ske en meget kraftig reduktion i den lokale support, hvis de totale supportomkostninger i Scenarium 2 skal blive lavere end i Scenarium 1*

6) Finansieringsmodel og forventede fremtidige omkostninger for investeringer, som tillader performance at stige i takt med den forventede internationale udvikling:

Vedrørende fremtidige omkostninger bemærker vi: Under de her opstillede forudsætninger om samme totale omkostninger er de fremtidige omkostninger per definition næsten de samme. Som vist under punkt 4) er dog såvel den totale performance som den totale nytteværdig, som dette svarer til, dog betydeligt mindre i Scenarium 2.

Vedrørende finansieringsmodel bemærker vi: Under begge modeller kan der forventes, fortsat at komme bidrag til den lokale kapacitet fra lokale bevillingshavere, som har meget specifikke behov knyttede til sine eksterne bevillinger. Det vil f.eks. være grundforskningscentre, og andre store bevillingshavere.

> **Scenarium 1:** Finansieringsmodellen forventes at være den, at de lokale universiteter overtager ansvaret for de lokale anlæg, herunder at afholde omkostninger for reinvestering, datacenter-udgifter, systemvedligehold og support. Fremtidige omkostninger forventes, iflg. DeICs tabel, at ligge på omkring 50 Mkr/år.
>
> **Scenarium 2:** Finansieringsmodellen forventes at være, at universiteterne dels betaler for regnekapacitet på det nationale anlæg, dels fortsætter at drive de lokale centre. Med fastholdte totale omkostninger = 50 Mkr/år vil der, efter fradrag af 23 Mkr/år i udgifter til det nationale Tier-1 anlæg, være 27 Mkr tilbage til de lokale centre.
>
> I dette scenarium forventes universiteterne at bede sine forskere om, via ansøgninger til råd og fonder, direkte eller indirekte at betale for regnetid ved det nationale center. De enkelte forskere vil i den situation sandsynligvis finde det mere givende, at søge om fri regnekapacitet direkte hos PRACE/DECI på Europæisk niveau, hvor

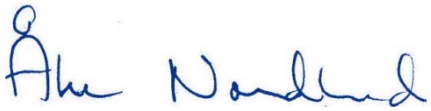
der er adgang til betydeligt større anlæg, og hvor ansøgningsmekanismen er fair og transparent; ansøgninger om regnekapacitet konkurrerer med hinanden, og ikke med ansøgninger om helt andre ting.

Der skal i denne forbindelse yderlige bemærkes, at den fremtidige finansiering for PRACE ressourcer i øjeblikket er under diskussion. HPC/KU er repræsenteret i arbejdsgruppen, som har til opgave at levere (kun) to modeller; en baseret på nationale kontantbidrag og en baseret på EU kontantbidrag. På intet tidspunkt er *in kind* bidrag kommet i betragtning for den fremtidige finansiering af PRACE. Fravær af et nationalt dansk Tier-1 anlæg vil derfor ikke, som der antydes i indlægget fra SDU, tælle negativt i samarbejdet omkring PRACE.

Sammenfattende konkluderer vi:

1) At **tidsrammen for installation** er mere fordelagtig i Scenarium 1
2) At de **initiale omkostninger,** som konsekvens af sammenligningens forudsætninger om samme totalomkostninger, er praktisk taget de samme
3) At **dækning af spektrum af behov** er mere fordelagtig i Scenarium 1 med hensyn til Tier-2 type behov, og at Tier-1 og Tier-0 type behov dækkes bedre, og uden nye udgifter, via det danske medlemskab af PRACE/DECI.
4) At **forventet performance medio 2014** vil være større i Scenarium 1, og at den forventede performance for det af SDU skitserede Tier-1 anlæg kun vil være ca. 1/3 af det anførte, da der skal tages højde for det reducerede antal GPGPUer.
5) At de **forventede fremtidige omkostninger for drift** vil være større i Scenarium 2
6) At de **forventede fremtidige omkostninger for investeringer**, som resulterer i sammenlignelig nytteværdi, vil være større i Scenarium 2

Med venlig hilsen,



Prof. Åke Nordlund, Niels Bohr Institutet, Københavns Universitet
(DCSC bevillingshaver 2003-2011, DECI og PRACE bevillingshaver 2009-2013)

Postdoc Troels Haugbølle, Statens Naturhistoriske Museum, Københavns Universitet
(PRACE bevillingshaver 2013-2014)