

Til: FOR-ANS Notur referansegruppe
Fra: FOR-ANS Notur teknisk gruppe
Forfatter: Jørn Amundsen og Roy Dragseth
Kopi: Arild Halsetrønning
Dato: 11.8.2015

Gjelder: Leveranser fra FOR-ANS Notur teknisk arbeidsgruppe

1. Innledning

Dette notatet beskriver leveransene fra FOR-ANS Notur teknisk arbeidsgruppe. Arbeidsgruppens mandat har to leveranser:

- Dokument med teknisk kravspesifikasjon basert på brukerbehovene, til bruk i punkt to og til bruk i anskaffelsesprosjektet.
- Dokument som beskriver alternative utbyggings- og utfasingsstrategier, med ulike installasjoner og med en anbefaling.

De to leveransene er beskrevet i kapittel to og tre nedenfor.

2. Krav til teknologiske løsninger

Arbeidet kan grovt deles i følgende deler:

1. Gjennomgang av resultat fra behovsundersøkelsen som ble gjennomført våren 2015.
2. Analyse av applikasjonsporteføljen i Notur.
3. Analyse av nåværende bruksprofil på Noturs regneanlegg.
4. Leverandørmøte.
5. Innsamling og gjennomgang av anbudsdokumenter benyttet ved tidligere anskaffelser i Norge, samt to fra Sveriges siste anskaffelser.

Behovsundersøkelsen ble gjennomført av Sigma2 som en nettundersøkelse, samt at det ble gjennomført besøk med brukermøter hos partnerne i Notur-prosjektet. Tilbakemeldingene fra behovsundersøkelsen og besøksrunden er at norske forskere i all hovedsak kan få sitt behov for regnekraft dekket av en utvidelse og fornyelse av de teknologiske plattformer som tilbys i Notur i dag. Det ble gitt tilbakemelding om at brukeroppgivelsen for dem som har behov for å teste og utvikle parallelle applikasjoner er for dårlig, og at det må tas grep for å forbedre dette.

I rapporten fra akseleratorprosjektet er det en grundig analyse av dagens applikasjonsportefølje som videre underbygger at det ikke vil være formålstjenlig nå med store investeringer i nye teknologier som f.eks. akseleratorer basert på NVIDIA GPU-er eller Xeon Phi. Det er imidlertid viktig for Notur å understøtte utvikler- og applikasjonskompetanse på denne typen arkitekturer, så det bør gjøres mindre investeringer innen dette området også. Rapportens uttaksstatistikk viser at det er kun et fåtall av de mest ressurskrevende applikasjonene som skalerer over 1000 kjerner. To av miljøene som benytter de mest ressurskrevende applikasjonene er svært synlige i sine fagfelt internasjonalt - klima og astrofysikk - så det er viktig at fremtidige regneressurser dekker disse miljøenes behov for regnekraft ved å ta høyde for at de valgte teknologiene understøtter en forbedring av applikasjonenes skalerbarhet over tid. Hvis man ser på lastgrafene for dagens Notur-anlegg (<http://www.notur.no/hardware/status/>) er bruksprofilen todelt mellom Hexagon/Vilje og Abel/Stallo. Hexagon/Vilje har en overvekt av jobber som kjører opp til ca. 4000 kjerner, mens hovedvekten av jobber på Abel/Stallo ligger under 512 kjerner. Dette gjenspeiles også i applikasjonsanalysen som er gjort i akseleratorprosjektet, hvor de mest ressurskrevende applikasjonene på Hexagon/Vilje er fra klima, astrofysikk og molekylodynamikk, og på Abel/Stallo fra kjemi og bioinformatikk. Bruksstatistikken indikerer også at hovedparten av brukerbehovene kan dekkes av beregningsnoder som har ca. 2 GiB minne per kjerne, dvs. standard serverteknologi. Det er også hentet ut statistikk for utnyttelsen av storminnenodene på Abel, åtte noder med 1 TiB minne hver, og det viser seg at over 12 måneder er det kun et fåtall tilfeller der mer enn 50% av minnet på en node har blitt benyttet. Konklusjonen er at det er et behov for denne typen tjeneste, men at det kan dekkes ved anskaffelse av et

fåttall storminnenoder, f.eks. 2-4 noder med minst 5 TiB, og et noe større antall noder med opptil 512 GiB minne per node.

Den 11. juni 2015 ble det gjennomført et leverandørmøte med seks leverandører på Gardermoen. Disse var Cray, SGI, Lenovo, Megware, Intel og HP, og var valgt ut fra at de hadde leveranser til Notur i dag og/eller er viktige premissleverandører for den videre teknologiutviklingen. Siden alle møtene var under NDA kan ikke detaljene beskrives her, men hovedpunktene er:

- Linux som operativsystem er fortsatt den dominerende plattformen
- Standard prosessorarkitektur med x86_64 er basis for alle plattformer, men alle har tilbud innenfor akselerorteknologier.
- De fleste leverandørene har nå vannkjøling som forutsetning for sine mest tettpakkede systemer.
- Nye systemer avgir mer varme per volumenhet og har høyere vekt per flateenhet.
- Det kommer et teknologiskifte i løpet av 2016 som gjør Xeon Phi til en selvstendig prosessorplattform.
- Det kommer en ny nettverksteknologi som vil konkurrere med Infiniband.

Teknisk kravspesifikasjon for ny anbudsprosess

Arbeidet med ny kravspesifikasjon ble innledet med en gjennomgang av anbudsdokumenter fra tidligere anskaffelser av tungregneutstyr i Norge og to svenske anbudsprosesser som ble gjennomført i 2014 av KTH og NSC. Etter gjennomgangen ble anbudsdokumentene fra UiO for innkjøpet av Abel valgt som grunnlag for det videre arbeidet, med en teknisk kravspesifikasjon og elementer fra de andre dokumentene innarbeidet i dette. Dokumentet ble omarbeidet til å passe med konklusjonene man trakk fra arbeidet referert ovenfor, med følgende hovedtrekk:

- Hoveddelen av systemet må være basert på x86_64 prosessorer med Linux som OS og ca. 2 GiB minne per kjerne.
- Et antall noder av samme type som over med 512 GiB minne.
- Lite antall noder med 5 TiB minne.
- Interkonnekt må ha lav latens (Infiniband eller tilsvarende).
- Globalt filsystem med støtte for valgt interkonnekt.
- Lite antall noder med akselerorteknologi.

Dokumentet er ment å danne et utgangspunkt for kravspesifikasjon i en anbudsprosess, men mange detaljer må utsettes inntil man har bestemt utbyggingsstrategi og det er bestemt i hvilket datasenter anlegget skal installeres.

3. Evaluering av utbyggings- og utfasingsstrategi

Evalueringen ble gjennomført i et todagers arbeidsgruppemøte på nøytralt sted. Den er basert på et forslag til metodikk fra UNINETT Sigma2. Sigma2 gav innspill til kriterier, basert på tidligere diskusjoner og elementer fra Dæhlen-utvalget, Forskningsrådets rapport «Nasjonal e-infrastruktur for forskning og høyere utdanning» (R. Skålin, 2013) og Sigmas erfaringsrapport. Mesteparten av tiden ble brukt til en grundig gjennomgang av kriteriene infrastrukturen må og/eller bør oppfylle. Man etablerte et sett av hovedkriterier, og innenfor hvert hovedkriterium et sett av delkriterier med sentrale elementer å oppfylle.

Etter at man hadde etablert konsensus om kriteriene, diskuterte man seg frem til relevante målarkitekturer.

Rangeringen ble utført helt til slutt ved at man først vektet hovedkriteriene og deretter vektet alle delkriteriene innenfor hvert hovedkriterium. Så ble det for hver arkitektur poengsatt på en skala 0-10 i hvilken grad den oppfylte hvert enkelt delkriterium.

3.1. Hovedkriterier

Arbeidsgruppen definerte seks hovedkriterier som en målarkitektur må oppfylle. Hovedkriteriene og de viktigste delkriteriene er, sortert etter fallende vekt:

Vekt	Hovedkriterium	Viktigste delkriterier
0.30	Dekning av brukerbehov	parallellitet innenfor ett rack, parallellitet > 1000 kjerner, parallellitet innenfor en node, båndbreddeintensiv I/O, standard minne/kjerne (2 GiB) - og knytning mot fagområder
0.25	Kostnadseffektivitet og driftssikkerhet	driftssikkerhet (resilience), investeringskostnad, driftskostnad, utnyttelse, utfasing og utbygging, innkjøpsbehov
0.15	Brukerorientering	muligheten for å frigjøre ressurser til avansert brukerstøtte, tilpasning av maskinressurs til fagdisiplin
0.15	Metasenter samarbeid	basis brukerstøtte, kompetanse, standardisering, prosjektsamarbeid, driftsmodell
0.10	Integrasjon	integrasjon Notur-NorStore, dynamisk ressursallokering (tilgang via portaler)
0.05	Spesielle behov	rask tilgang til regneressurser, undervisning/utvikling/testing/prekvalifisering, lokale strategiske hensyn

For hvert delkriterium er det fylt inn viktige elementer å oppnå. Se vedlagt regneark for fullstendig informasjon.

Krav til diskvolum er ikke evaluert for noen av alternativene, siden det ikke har konsekvens for valget av målarkitektur. Dette må vurderes på en senere punkt i anbudsprosessen. Lagring og prosessering av sensitive data (TSD) er heller ikke evaluert, fordi det må være en relativt liten maskin hvor maskinvare ikke kan integreres med øvrig utstyr. Arbeidsgruppen ser det mest formålstjenlig at TSD leveres med utgangspunkt i dagens leveranse fra USIT, men rolle- og tjenestedefinisjon må avklares.

3.2. Evaluerte målarkitekturer

Arbeidsgruppen diskuterte flere mulige konfigurasjoner, men anså tre målarkitekturer som relevant og tilstrekkelig for evalueringen:

1. Alternativ S+L: En (stor) HPC-maskin, inkludert en lavterskelmaskin.
2. Alternativ 2S+L: To HPC-maskiner (A1 og B1, ev. profilert med forskjellig applikasjonsportefølje), og en lavterskelmaskin.
3. Alternativ 4M: Dagens løsning med en maskin ved hvert av de fire metasenteruniversitetene.

Med begrepet *maskin* menes et sett av mulig heterogene noder i ett eller flere rack, men i samme maskinrom og med et felles, hurtig, internkommunikasjonsnettverk (interkonnekt). I alternativ S+L ser man for seg at en maskin skal betjene hele Notur-lasten, mens den i alternativ 2S+L er tenkt fordelt over to maskiner, med mulig forskjellig applikasjonsportefølje.

Lavterskelmaskinen L er et mindre system med samme maskin- og programvare som HPC-maskinene S. Hensikten med en lavterskelmaskin er å fasilitere programutvikling, opplæring og en interaktiv omgivelse. Da kan nye brukere få en mykere start før overgang til HPC-maskinene. Det kan også være mulig å få lavterskelmaskinen levert tidligere ved en anbudsprosess, slik at man kan komme i gang med forberedelse av programvare før de store maskinene kan installeres. Det er en forutsetning at lavterskelmaskinen plasseres maskinromsnært til HPC-maskinen, eller ved en av dem i alternativ 2S+L.

Alternativ 4M er en videreføring av dagens regime (som målarkitektur, men ikke utbyggingsstrategi).

3.3. Resultat av evalueringen

Etter fastsettelse av vektorer for alle kriterier endte evalueringen med ni av ti poeng for alternativ 2S+L, åtte av ti for alternativ S+L og fem av ti for alternativ 4M. Resultatet av poengsettingen er oppsummert i tabellen nedenfor:

Hovedkriterium	Uttelling (0-10) for hvert alternativ		
	1	2	3
Dekning av brukerbehov	10	10	9
Metasenter samarbeid	10	10	2
Kostnadseffektivitet og driftssikkerhet	5	9	6
Spesielle behov	9	10	8
Brukerorientering	8	10	4
Integrasjon	10	6	1
SAMLET VURDERING	8	9	5

I motsetning til alternativ 4M ble alternativene S+L og 2S+L vurdert som nesten like gode. Momenter som har gitt poengtap i evalueringen er:

- Alternativ 4M får dårlig uttelling på metasenter samarbeid fordi det av erfaring har vist seg vanskelig å etablere felles standarder og kjøre felles prosjekter på en effektiv måte i konkurranse med lokalt arbeid.
- Alternativ S+L får dårlig uttelling på kostnadseffektivitet og driftssikkerhet, fordi gruppen vektlegger driftssikkerhet som avgjørende for tjenestens brukbarhet.
- Alternativ 4M har god driftssikkerhet, men scorer gjennomgående dårlig på delkriteriene investeringskostnad, driftskostnad, utnyttelse og ressurskrevende innkjøpsbehov.
- Alternativ 4M scorer dårlig på brukerorientering fordi det ansees bemanningskrevende og gir lite rom for frigjøring av ressurser til avansert brukerstøtte.
- Alternativ S+L scorer klart best på integrasjon (NorStore og portaler), mens alternativ 2S+L har fått dårligere uttelling fordi kun en HPC-maskin kan få en høykvalitet Notur NorStore-integrasjon dersom de to HPC-maskinene plasseres geografisk adskilt.
- Alternativ 4M gjør en god integrasjon mot NorStore svært vanskelig, og er det mest utfordrende alternativet å integrere i portalbruk.

Arbeidsgruppen mener at man ved alternativ S+L kan oppleve at hele beregningstjenesten er utilgjengelig ved planlagte eller ikke-planlagte driftsavbrudd. Det ble vurdert som ikke akseptabelt.

Arbeidsgruppen anbefaler alternativ 2S+L, to HPC-maskiner med en lavterskelmaskin.

3.4. Forslag til utbyggings- og utfasingsstrategi

Forslag til utbyggings- og utfasingsstrategi er vist i fanen «Tidslinjer» i vedlagte regneark. Installasjonstidspunktene er angitt med en dobbeltlinje. Tidslinjen for den foreslåtte målarkitekturen 2S+L er:

- 2016: Økt behov dekkes ved utvidelse av Abel og/eller Stallo
- 2017: Installere HPC-maskin A1 og lavterskelmaskin L1, fase ut Hexagon og Vilje
- 2018: Installere HPC-maskin B1, fase ut Abel og Stallo
- 2019: Nødvendig økning av kapasitet etter behov på A1 og B1
- 2020: Installere maskiner A2+L2, fase ut A1+L1
- 2021: -
- 2022: Installere HPC-maskin B2, B1 fases ut

Alternativ 2S+L har en transient i installasjonsforløpet ved at B1 anskaffes allerede året etter A1, for hurtigst mulig å fase ut Abel og Stallo. Dette kan justeres eller utsettes, avhengig av finansiering. I alternativene S+L og 2S+L etableres det ett felles driftsteam med deltagelse fra alle

metasenteruniversitetene. Det bør være en trinnvis migrering av driftsteamet fra dagens maskinvare til den nye infrastrukturen.

Ved vurdering av alternativ 4M har man valgt å ikke videreføre samtidig fornyelse av alle maskinene, men å trinnvis erstatte ett system hvert år for å unngå dårlig kapasitetsutnyttelse. Alternativet vil gi fire relativt små systemer, som pga. en annen finansieringsmodell vil bli mindre enn dagens maskiner.