SPC/3 PROJEKTET

SUPERMAX VIDEREUDVIKLING

Referencer:

1. SPC/3 Rapport over forprojekt. BW-aug.91.
2. Supermax som Multiserver i Client/Server arkitekturer. FL/SST-2.10.91.
3. DDE 'Road Map'. Draft, CT-okt.91.

1. INTRODUKTION

Denne rapport skitserer en model for videreudvikling af Supermax datamaten. Modellen realiserer nogle nødvendige generationsskift ved at bygge videre på de erfaringer, der ligger i den nuværende Supermax.

Rapporten består først af en gennemgang af de **forudsætninger**, der ligger til grund for planen. Dvs. hvilke miljøer vi forventer at sælge maskiner i, og hvilken udvikling vi forventer vil finde sted, markedsmæssig og teknologisk. Dette munder ud i en beskrivelse af de overordnede designkriterier.

Derefter følger et afsnit om 'nøglevalg', dvs. de afgørende valg, som har væsentlig indflydelse på Supermax's egenskaber fremover, samt et afsnit om en foreslået produktstrategi. Disse 2 afsnit beskriver således, hvorledes de ovennævnte designkriterier kan opfyldes.

Efterfølgende er indlagt en skitse til implementation, som i mere tekniske detajler viser, hvorledes maskinen kan opbygges. Dette afsnit gengiver også argumentationen for valgene i I/O-arkitekturen.

Endelig er der et afsnit om økonomi og en skitse til ressource og tidsplan.

Rapporten er blevet til på grundlag af den tidligere udsendte rapport om SPC/3 projektet (Ref.1), den efterfølgende diskussion med direktionen, samt det senere arbejde i projektgruppen.

2. FORUDSÆTNINGER

Supermax indgår idag i en række anvendelser og miljøer, som på udmærket vis er beskrevet i skriftet **Supermax som Multiserver i Client/Server arkitekturer** (Ref.2). Der er heri ligeledes skitseret en strategi for Supermax's rolle i Client/Server miljøer. Denne strategi kan kort resumeres med følgende:

Supermax skal udnytte de fordele der ligger i Client/Server arkitekturen. Dette gøres ved at sikre at Supermax har de rigtige karakteristika og de rigtige værktøjer, herunder specielt de rigtige kommunikationsmekanismer til at gøre den specielt velegnet som server. Opgaver, som bedst placeres i centrale servere, skal kunne afvikles optimalt på en Supermax.

De afledede krav af ovenstående dækker områder som:

- åbent system
- sikkerhed
- robusthed
- administration og overvågning
- skalering og modularitet.

Den nuværende Supermax skal videreudvikles med ovenstående for øje, idet der idag er nogle af ovenstående punkter, hvor

Supermax ikke er eller vil være tilstrækkelig dækkende (se nærmere herom i afsnit 2.7).

Først nogle overvejelser over hvilke elementer, der kan tænkes at have indflydelse på arkitekturen i Supermax videreudviklingen.

2.1. Miljøer

De (tekniske) miljøer Supermax indgår i er ofte meget blandede. Men hvis de 'rendyrkes', kan disse miljøer deles i 3 forskellige grupper med klart adskilte karakteristika:

1) PC-net, Novell eller Lan Manager baseret. Der opereres med MicroSoft's Client/Server princip med database back-end servere, med trafikken over nettet så vidt muligt begrænset til SQL-kommandoer. PC'erne kører typisk MS-Windows, og LAN Manager afvikles enten på en OS/2 baseret PC-server, eller en UNIX maskine med LM/X. Der er ingen terminaler. Lan Manager konfigurationen er et af de tilfælde, hvor OS/2 og UNIX operativsystemerne er i direkte konkurrence. Da LAN Manager oprindelig er skrevet til OS/2 (begge dele udviklet af MicroSoft), er den idag mere velkørende på OS/2 end på UNIX (LM/X).

   Supermax's styrke som **LM/X-server** og **database back-end server** må her ligge i den modulære opskalering til en overlegen ydeevne, såvel CPU-mæssig, som Disk/Net I/O.

2) Workstation miljø med forbundne UNIX baserede workstation fra SUN,IBM,DEC, HP/Apollo m.fl. Der er typisk tale om teknisk baserede anvendelser som udgangspunkt. Der er i de senere år sket en kraftig vækst i anvendelsen af disse miljøer, således at de idag også anvendes til andre løsninger. Workstation teknologien har på en række områder bidraget med innovativ teknologi, idet den dedikerede grafik, UNIX operativsystemet med tilhørende LAN kommunikation, og de kraftigere RISC baserede workstations har åbnet op for nye anvendelser og nye løsninger. Således er håndteringen af billeder og farver lettere på workstations end på PC'er. Se iøvrigt afsnittet om multimedie.

   I disse miljøer er der behov for en filserver funktionalitet i stil med PC-nettets, idet man ofte vil lægge større mængder fælles data på en **NFS-server**. Det er denne funktionalitet f.eks. SUN's nye SPARCserver 600 serie retter sig mod. Kravene til maskinarkitektur er i stor udstrækning de samme som under 1).

3) Central UNIX maskine, anvendt som data-base server og applikation-server. Terminalerne vil efterhånden udelukkende være tilsluttet lokalnet, og være X-Windows baserede.

Kravet om X-Windows baserede løsninger kommer først på grafiske applikationer, fordi det er forudsætningen for at gøre løsningerne portable. På ikke-grafiske applikationer vil de MS-Windows baserede alternativer på PC løsninger skabe et ønske om tilsvarende brugerinterfacer på UNIX-maskiner. Dette resulterer i anvendelse af OSF/Motif (som forudsætter X-Windows).

Dette er naturligvis Supermax's 'hjemmemiljø'. De løsninger der er specielt velegnede til dette miljø er ofte karakteriseret ved at have (store) centrale databaser. Brugen af X-terminaler og grafisk brugerinterface øger kravene pr. bruger til LAN I/O og til CPU-ydelse.

Man må forvente at ovenstående 3 katagorier vil danne grundlag for de fleste løsninger i fremtiden. Naturligvis vil man ofte se konfigurationer, hvor ovenstående katagorier er blandede, således f.eks at X-Windows emulering foregår på PC'er i et PC-Net, hvorved UNIX maskinen afvikler dels centrale X-Windows baserede applikationer, dels fungerer som LM/X server. Tilsvarende kan workstation miljøet tænkes at have behov for en databaseløsning som samtidigt afvikles på NFS serveren.

I alle de ovenstående miljøer kan serveren have kommunikations-linier til andre maskiner, og derved samtidig fungere som kommunikationsserver.

Som beskrevet i Ref.2 om Supermax som Multiserver skal vi sikre, at Supermax er optimal i disse serversammenhænge. Arkitektur-mæssigt sætter dette specielt krav til I/O båndbredde, se afsnit 3.2.

2.2. Antal brugere, sikkerhed og fejltolerance.

Som det fremgår af ovenstående, skal man ikke forvente nogen klar sammenhæng mellem ydeevne og antallet af brugere pr. system eller pr. maskine. Antallet af brugere, der kan kobles på en Supermax, er afhængig af anvendelsen.

De største nuværende installationer er på ca 250 brugere. Muligheden for at sætte flere brugere på en Supermax kommer direkte af den teknologisk betingede forøgelse af CPU-ydelse, forudsat uændret applikation. Allerede de nuværende største konfigurationer kræver en række overvejelser omkring, hvorledes sådanne 'Main-frame miljøer' skal håndteres. Supermax skal kunne leve op til disse krav vedrørende sikkerhed, fejltolerance og værktøjer til system administration.

Som nævnt i SPC/3 rapporten (Ref.1) vil vi med fejltolerance forstå 'high availability', som omfatter:

Robusthed:
1. Power backup på disk og disk cache (jvf. Gogolplex i Ref.1).
2. Data beskyttes og checkes i hele systemet, så fejl opdages på et så tidligt tidspunkt som muligt, og fejlen rettes om muligt (parity check og ECC).

Dublering:
1. Mulighed for redundante systemer som f.eks. Supermax Dual Hosted Disk System.
2. Anvendelse af spejlede diske (evt. alternativ teknik), kombineret med udskiftning af diske under drift.

Kort 'down-tid':
1. Hurtig fejllokalisering og moduludskiftning.
2. Kort boot-tid.

Generelt vil vi gennemføre en undersøgelse af hvilke afledede krav og forventninger man i Main-frame miljøer har til sikkerhed, fejltolerance og system administration, samt en analyse af, hvordan man yderligere kan styrke ovennævnte maskiner mod høj tilgængelighed.

Om sådanne 500-1000 bruger kunder eksisterer for DDE er et et uafklaret spørgsmål.

2.3. Skalerbarhed og modularitet.

Supermax skal bruges i mange koncepter, hvor der er vidt forskellige krav til maskinens ydelse, kost og krav til maskinens særegne kompetance. Den bedste måde at imødekomme den form for krav, ligesom imødekommelse af de ovenfor nævnte server-ønsker, er at udvikle en maskine med en høj grad af skalerbarhed og modularitet.

Den har derved en styrke i den brede dækning af performance spektret og i serverspecialisering. Denne styrke skal kunne kompensere for den manglende konceptspecialicering.

Denne manglende konceptspecialisering ses f.eks. i ECAD-løsningen. En minimum konfiguration tilkoblet en enkelt X-terminal udgør som enkeltstående maskine ikke umiddelbart et alternativ til en dedikeret workstation. Kun hvis produktets skalerbarhed og modularitet dækker kundens behov i højere grad end den dedikerede workstation, er der tale om et reelt alternativ. Se iøvrigt afsnit 3.4 om ACE-PC'er.

2.4. Åbne systemer og bagudkompatibilitet.

Det er DDE's strategi at overholde en række standarder vedrørende "-bility" funktionalitet (jvf. Ref.2):

- Compatibility:
  Under dette punkt hører, at vi skal kunne videreføre vore kunders investeringer i software ved hjælp af bagudkompatibilitet, og **samtidig** at vi ønsker, at Supermax skal placere sig i en verden af binær kompatibilitet.

- Portability:
  Under dette punkt hører, at vi ønsker, at standard applikationer fra andre UNIX verdener skal kunne porteres til Supermax.

- Interoperability
  Herunder hører, at vi ønsker at kunne anvende samme kommunikationsstandarder som andre tilsvarende systemer.

- Scalability:
  Herunder hører, at programmellet på alle niveauer skal kunne være identisk fra de helt små til de helt store maskiner.

To konsekvenser af ovenstående:

1) Konflikten i Compatibility-punktet kan tvinge DDE ud i et brud med bagudkompatibiliteten (se afsnit 3.3 vedr. OS-valg og little/big endian). Dette er tidligere afhjulpet ved hjælp af de såkaldte 'heterogene' maskiner.

2) De første 3 punkter skaber et krav til sammenhæng med den øvrige verden omkring os. Derfor anbefales, at styresystemet på Supermax med tiden bliver et standardstyresystem, der så vidt muligt skal kunne portes med en minimal tids- og ressourceindsats. Der findes også andre grunde til dette ønske, se afsnit 3.3 vedrørende *teknologiresponstid*.

2.5. Multimedie

Det er i workstation miljøet vi først vil se anvendelse af multimedie, med ønske om håndtering af audio, still-billeder og video, med deraf følgende krav til beregningstunge algoritmer til komprimering mm., og store krav til lagring af data.

En workstation med videokobling kan således afspille et videobånd i et vindue på skærmen, eller indspille på video en sekvens af begivenheder som finder sted i et vindue.

Man må forvente, at multimedie i lang tid primært vil være knyttet til workstations, og serverens rolle (og dermed Supermax's rolle) er modtagning, lagring og sending af store (komprimerede) datamængder. Dette vil kræve stor båndbredde i systemet.

2.6. Konkurrenter.

Indtil nu har det været forbeholdt de få at kunne levere UNIX multi-CPU maskiner baseret på standard microprocessorer, der anvender delt lager.

Således har Sequent en Intel baseret multi-CPU maskiner, og hævder at have leveret 4000 systemer. Pyramid og Encore har ligeledes begge et UNIX multiprocessor system.

Dette forhold er nu ved at ændre sig.

For det første findes der nu en multiprocessor version af SCO-UNIX til Intel baserede maskiner. Denne kan anvendes på Compaq's dobbelt-CPU system. Den almindelige tilgængelighed af dette system via SCO gør, at mange leverandører nu er i stand til

at udvikle deres egen Intel baserede multi-CPU maskiner, uden at skulle tænke på problemer med at få operativ systemet til at fungere. Det er det ICL har gjort med deres 4x486 maskiner. Versionen er leveret af Corollary og er ikke en fuld symmetrisk UNIX, hvad der begrænser antallet af CPU væsentligt.

For det andet er USL ved at færdiggøre en fuld symmetrisk multi-processor version af SVR4, kaldet SVR4/MP. Denne er udviklet i samarbejde med nogle af de væsentlige UI medlemmer f.eks. NCR og Olivetti, og findes nu i en test version til Intel. Denne udvikling gør, at i princippet alle vil være i stand til at levere en sådan UNIX multi-processor maskine.

Eksempel:

SUN har annonceret SPARCserver 600MP serien med op til 90 SPECmark på en 4-CPU maskine med mulighed for tilslutning af 26 GB disk. Maskinen er annonceret med en ikke-symmetrisk multiprocessor SUN-OS. SVR4/MP med fuld symmetrisk multiprocessing forventes næste år. Det forventes ligeledes, at SUN næste år vil introducere en kraftigere maskine baseret på Texas Instruments superscalar SPARC multiprocessor med kodenavnet 'Viking'.

SUN's maskine er ikke et back-plane baseret system, hvad der begrænser antallet af CPU'er til nogle få. Det er formentlig det, de fleste vil gøre i starten, ikke mindst på workstations. SVR4/MP muliggør en symmetrisk sammenkobling af overordentlig mange homogent og symmetrisk sammenkoblede CPU'er, og disse vil vi derfor også se som backplane maskiner. Imidlertid understøtter SVR4/MP ikke endnu Supermax-lignende arkitekturer, ligesom det vil være de færreste, der som DDE har egentlig erfaring i optimaliseringen af dem.

DDE står derfor bedre rustet end de fleste andre i udviklingen af konkurrencedygtige maskinarkitekturer til support af multi-CPU UNIX.

Andre eksempler:

HP's workstations (1 CPU) yder idag op til 72 SPECthruput. Med deres tidligere RISC processor har de en 4-CPU maskine.

DEC forventes at ville bygge maskiner med 8 MIPS-CPU'er. Silicon Graphics har en sådan. MIPS har cancel'et deres projekt. CDC har baseret deres Midrange 4300 serie 'Infoservers' på teknologi fra Silicon Graphics, og de har en højere ydende 4680 serie baseret på MIPS R6000 med op til 205 SPECthruput med teknologi fra MIPS.

NCR har formentlig en del erfaring fra deres multiprocessor Tower systemer, men de anvender ikke delt lager. Deres Intel baserede 4-8 CPU systemer anvender SCO/Corollary og vil anvende SVR4/MP fra USL.

Olivetti har demonstreret sin PWS4000 baseret på MIPS R4000 single CPU. Bull er med i 'Mustang' projektet, hvor LSI udvikler kredse til support af optil 4 stk. R4000 fra MIPS. SNI anvender

også MIPS, men vi kender ikke deres planer.

IBM arbejder med sammenkobling af deres RS6000 boxe med 220 Mbit/s optiske links, og har udviklet multiprocessor software til en 'Cluster Manager' og en 'Distributed Lock Manager' til koordinering af løst koblede RS6000 maskiner bundet sammen med disse links. De planlægger iøvrigt at udvikle deres egen symmetriske multiprocessor arkitektur.

Flere leverandører arbejder med maskiner med sigte mod andre arkitekturer. Det handler om large scale multiprocessing baseret på parallel processing uden anvendelse af delt lager. Der foregår meget research på området, men indtil videre er disse maskiner dedikerede i den forstand at de kræver specieltudviklet applikationssoftware.

I den lave ende kan vi forvente konkurrence fra ACE-PC'er baseret på MIPS CPU. Hvad dette er, har SGI givet en forsmag på med deres Indigo workstation. Det er en R3000 baseret 'ACE-compliant' maskine med 8MByte RAM, 236 MByte disc, 24bit farver emuleret til 8bit på 16 inch monitor, med UNIX, Motif, 3D grafik mm. for $9995. Inkluderet er tilslutning af video og audio. Den er interessant, fordi den sætter et nyt niveau for hvad man som minimum kan forlange at få i en workstation af datakraft, grafik og multimedie for under $10.000. Sådanne dedikerede maskiner er det naturligvis ikke rimeligt umiddelbart at sammenligne et modulært og skalerbart produkt som Supermax med. Men vi er nødt til at tage stilling til hvordan vi skal håndtere konkurrencen. Se afsnit 3.4 vedrørende ACE PC-er.

2.7. Supermax i dag.

Supermax er dimensioneret i begyndelsen af '80-erne. At dens fundamentale tekniske ramme har nået grænsen for, hvad man med begrænsede virkemidler kan forbedre, er ikke overraskende. Det overraskende er snarere, at det har været muligt at indlægge teknologisk opdaterede processorer og I/O komponenter i den eksisterende ramme over så lang en periode. Den har haft en forbløffende indbygget mulighed for 'skalering over tid'.

At denne grænse er nået, viser sig primært ved at modulariteten i dag kun giver begrænset skalerbarhed. Den kraftigste RISC maskine består reelt kun af 2-3 CPU'er. En Supermax kan reelt kun drive lokalnet op til 400 kbyte/s. Der er flaskehalse i systemet, både bus og I/O båndbredde. Der er således ikke længere den store forskel på et lille og et stort system.

En Supermax kan i dag udbygges ved at sammenkoble flere Supermax'er via lokalnet ved hjælp NFS o.lign. Dette giver til en vis grad en mulighed for at øge antallet af tilsluttede brugere (afhængig af applikationen). Som også nævnt i Ref.1 er der imidlertid en række fordele ved at anvende en maskine med en noget tættere kobling, som f.eks fælles processhierarki og bedre fordeling af CPU belastning.

Supermax skal med andre ord have en teknologisk opdatering af både bussystemet og I/O systemet, således at det er muligt at

opbygge store maskiner. Samtidig skal Supermax fortsat have udbygget sine muligheder for lokalnetsammenkobling ved hjælp af for eksempel DCE.

2.8. Designkriterier

Som konklusion på ovenstående overvejelser kan vi opstille en liste over grundlæggende designkriterier i Supermax videreudvikling:

- Skalerbarheden skal bringes væsentlig videre, behovet for datakraft vokser med muligheden for at tilbyde den. Kravet på performance går på CPU, Disk I/O båndbredde og Disk kapacitet, og LAN I/O båndbredde.

- Programmel kompatibilitet imellem alle modeller fra de mindste til de største.

- De større centrale maskiner stiller større krav til sikkerhed, robusthed, high availability, værktøjer til system administration, svarende til nuværende Main-frame installationer.

- Åbne systemer med deraf større krav til 'teknologiresponstid'.

- Med focus på skalerbarhed mod store systemer skal de helt små systemer ikke glemmes, men skal måske håndteres specielt.

Det primære sigte i arkitekturen og dermed styrke i Supermax er og skal stadig være en høj grad af skalerbarhed, fra små flerbrugermaskiner/servere til store installationer med mange hundrede brugere. Dette lyder muligvis som en selvfølge, men det er væsentligt at bevare focus på dette ved større teknologiskift.

Der skal et generationsskift til at sikre, at denne skalerbarhed kan bevares ved stor ydeevne. Men også til sikre at maskinen er 'skalerbar over tid', dvs. at det over en årrække til stadighed er muligt med lille tilføjelse at øge ydeevnen og kapaciteten, ved hjælp af ny teknologi.

Det er og skal fortsat være Supermax unikke 'selling point' og dens differentierende faktor, at den tillader en spændvidde i ydeevne/kapacitet med mange ens moduler. Dette giver igen kunden sikkerhed for at kunne vokse med sit behov.

3. NØGLEVALG

3.1. CPU-Arkitektur

Begreberne 'løs-kobling' og 'tæt-kobling' har vist sig uheldige i vores diskussioner om maskinarkitektur, idet nogle betegner den nuværende Supermax arkitektur som 'tæt-koblet', andre betegner den som 'løs-koblet'.

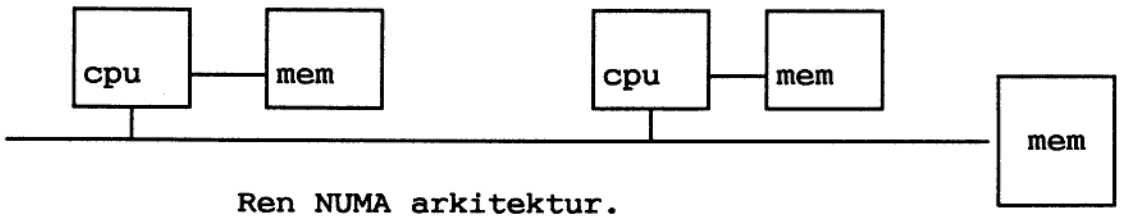
Disse betegnelser indgår derfor ikke i det følgende. Der benyttes kun betegnelser som er internationalt (nogenlunde) entydigt anvendt.

UMA = unified memory access, svarer til Trin 1 i Ref.1, dvs en symmetrisk multi-CPU, med et delt lager der har ens tilgang fra alle CPU'er overalt i addresserummet.

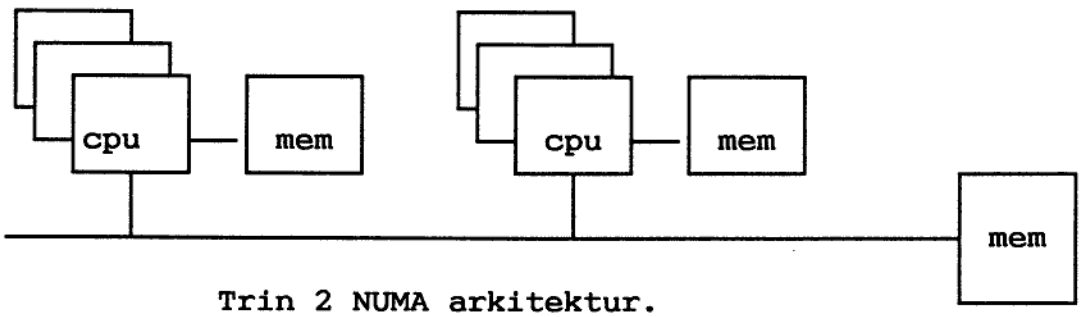
NUMA = non-unified memory access, svarer til en nuværende Supermax, med delt lager, hvor en del af addresserummet har en afvigende tilgang.

NORMA = no remote memory access, dvs der er ikke delt lager. CPU'erne udveksler information med en eller anden form for message passing. Det svarer til et system af Supermax knyttet sammen vha. NFS, RPC og lign. Det kan også være en arkitektur, der retter sig mod large scale multiprocessing, hvilket kræver specielt udviklet applikationssoftware. Disse arkitekturer overholder ikke kravene til 'åbne systemer'.

SPC/3 rapporten (Ref.1) beskriver nogle mulige arkitekturer for sammenkobling af et antal CPU'er ved hjælp af cache coherente busser. Trin 1 er en UMA arkitektur, mens Trin 2 er en speciel blanding af UMA og NUMA, idet det er en NUMA sammenkobling af nogle CPU-klynger, som selv er UMA sammenkoblet. Det interessante ved denne arkitektur er bl.a., at den i specialtilfældet hvor der kun er 1 CPU i hver klynge, har samme NUMA arkitektur som en Supermax idag. I det andet specialtilfælde er det en ren UMA maskine.



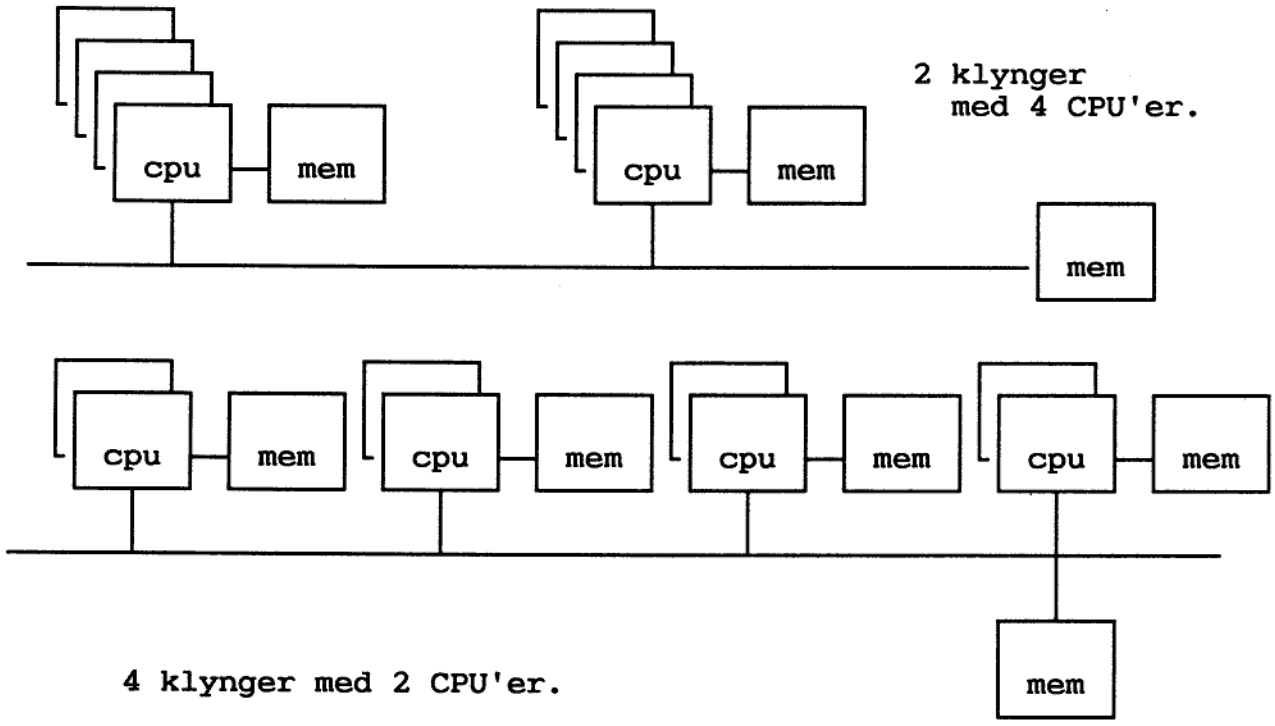
Ren NUMA arkitektur.



Trin 2 NUMA arkitektur.

Vi har valgt at sigte mod en sådan Trin 2 arkitektur.

For at sikre en skalerbarhed 'over tid' skal bussen have en båndbredde, der er på højde med den tilgængelige teknologi. Dette sætter en grænse for backplanets længde og dermed antallet af moduler. Vi har derfor valgt en 'lille' Trin 2 arkitektur med mulighed for tilslutning af i størrelsesordenen 16 CPU kort (afhængig af forbrug af slots til lager mm.). For at forenkle arkitekturen har de to busser samme (store) båndbredde. Disse 16 CPU'er kan konfigureres på de to busser, alt efter hvad der er mest optimalt. Det kan være 1 'klynge' med 16 CPU'er (ren UMA), det kan være 16 'klynger' med 1 CPU, eller et hvilket som helst mix. Dette gøres hardwaremæssigt ved hjælp af forskellige back-planes. Nedenfor er vist 2 eksempler på sådanne konfigureringer.



4 klynger med 2 CPU'er.

Baggrunden for dette valg er den erfaring, vi idag har med spredning af uafhængige processer over flere CPU'er i en Supermax. I Trin 2 vil processer, der arbejder uafhængigt (uanset om det er DDE software eller ej) kunne sprede sig over flere CPU klynger og på den måde udnytte den til rådighed værende ydelse.

Men nogle leverandører vil også udvikle software, der kan arbejde parallelt (og indbyrdes afhængigt) på sådanne Trin 2 NUMA arkitekturer.

ORACLE har med deres oracle 7 (også 6.2) lavet support af en multi-instance kerne. De enkelte instances har deres eget delte data område (svarende til nuværende SGA (=System global area)),

og de forskellige instances kommunikerer via en lock-manager, hvis services er arkitekturafhængig. På en Trin 2 vil de enkelte instances, som kunne tænkes fordelt på de forskellige UMA CPU-klynger, drage fordel af, at der er flere CPU-er til rådighed, da en enkelt instance består af flere processer. Disse processer kan arbejde på deres delte data uden at skulle ud på den globale fælles bus. Fælles bussen kommer først ind i billedet, når lock-manageren skal koordinerer låsninger mellem de forskelligge instances.

En Trin 2 Supermax er således meget ideel til fremtidige versioner af Oracle.

Mindst ligeså interessant er det dog at bemærke, at de store software leverandører, er begyndt at interessere sig for andet end de traditionelle symmetriske shared memory arkitekturer (dvs. UMA).

Der vil dog også være en modsat rettet tendens. SVR4/MP, som nu er almen tilgængelig, retter sig mod rene UMA maskiner. Dette vil give sig udslag i, at der vil komme flere softwareprodukter, som antager, at de befinder sig på en UMA multi-CPU maskine. De må formodes at ville supportere 'threads', der spreder sig over flere ens CPU'er.

Trin 2 NUMA maskinen har en usædvanlig arkitektur. Ikke fordi det er en arkitektur, hvor der til hver processor hører et lokalt lager, og hvor hver processor har et fælles globalt lager; det er som Supermax. Heller ikke fordi den tillader det lokale lager at blive delt mellem et antal processorer; det er blot en udvidelse af den lokale processor til en CPU-klynge. Det usædvanlige er snarere, at de 2 busser er ens og har samme båndbredde.

Dette skaber en flexibilitet og en usædvanlig mulighed for *tuning* af maskinen. Som antydet ovenfor er det meget applikationsafhængig, hvilken konfiguration der er optimal. En Trin 2 maskine kan konfigureres (og styresystemet tunes), således at belastningen fordeles bedst på de 2 busser, uanset hvordan applikationen er skruet sammen. Supermax vil således være optimal til afvikling af begge de to ovenfor nævnte typer applikationer.

Denne optimale belastningsfordeling vil forskydes med tiden, efterhånden som hastighedsforholdet mellem CPU og bus forskydes med stigende klokfrekvenser i CPU'en. Arkitekturens flexibilitet gør det enkelt at tage højde for denne forskydning, og der er derfor på forhånd indlagt en mulighed for skalering 'over tid'.

3.2. I/O Arkitektur

Nøglevalgene omkring I/O ser således ud:

1. Enhver maskine har et *Basalt I/O kort*. Dette indeholder SCSI, LAN/WAN, konsol og forskellige faciliteter krævet i ARC specificationerne. Dette er eneste I/O i små maskiner.

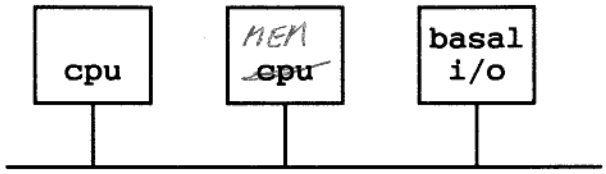
2. Der er specielle primære *I/O moduler* til Disk I/O og LAN I/O til at sikre I/O båndbredden. Disse I/O moduler kan sættes direkte på systembussen.

3. I større systemer er der en speciel *I/O bus*. Denne har samme specifikationer som systembussen. I/O modulerne kan derfor sættes på denne bus i systemer, hvor systembussen skal aflastes. I/O bussen forbindes til systembussen via en I/O bro i det globale lager.

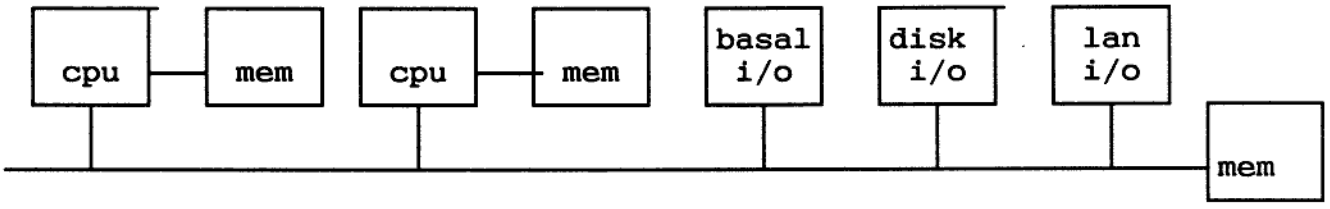
4. For at kunne sætte 'eksotiske' devicer på maskinen, kan man forbinde en *standard I/O bus* til systemet. Dette sker ved hjælp af en I/O bro, forbundet enten til I/O bussen eller til systembussen. Der er endnu ikke valgt standard bus. Det kunne f.eks være EISA.

Argumentationen bag ovenstående nøglevalg er ikke medtaget i dette afsnit, men er gennemgået i Afsnit 5: 'Skitse til Implementation'.

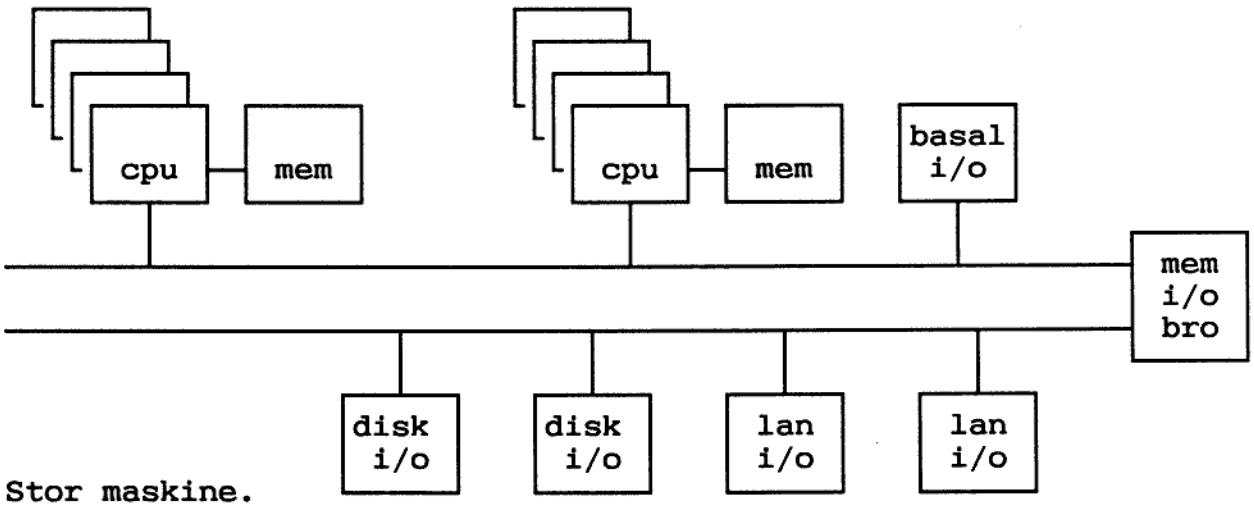
Vi kan herefter skitsere arkitekturen for tre forskellige størrelse maskiner, en lille, mellemstor og en stor:



Lille maskine.



Mellemstor maskine.



Stor maskine.

På denne måde kan man opbygge et meget stort spektrum af maskiner ved at benytte kun 6 forskellige moduler. Dette opnås bl.a. ved at benytte den samme type bus tre gange. Dels til de to systembusser, dels som I/O bus.

3.3. OS og ACE - og endians

Vi har i de seneste år set, at styresystemleverandørerne kommer hurtigere og hurtigere ud med nye features i deres styresystemer. Det er blevet mere og mere vanskeligt for DDE's styresystemudviklere at følge med denne udvikling, når alle nye faciliteter skal porteres til DDE's eget styresystem. Også i fremtiden ventes det at en stor strøm af nye faciliteter vil finde vej ind i styresystemerne, og hvis DDE skal have en rimelig kort "time to market" er det nødvendigt at kunne få disse nye faciliteter ind så hurtigt som muligt. Dette gøres enkelt og med få ressourcer ved at lægge en leverandørs styresystem næsten uændret ind på SPC/3. Denne evne til hurtigt og med få ressourcer at kunne lægge nye generelt tilgængelige faciliteter ind i systemet kalder vi en optimeret '**teknologiresponstid**'. De faciliteter der er tale om her er f.eks. NFS, RPC, DCE, og ikke mindst *enhanced security*.

Når og hvis ARC-maskinerne med deres tilhørende styresystemer slår igennem, vil vore kunder forvente at vort styresystem er funktionelt identisk med de andre ARC-maskiners. Dette kan vi naturligvis godt opnå ved en stor indsats af ressourcer på vort eget styresystem, men det vil være enklere at opnå denne ensartedhed med et standardstyresystem.

Vi anbefaler derfor at Supermax's styresystem med tiden bliver et standard styresystem. Dette skal naturligvis ikke forhindre, at vi foretager de ændringer/tilføjelser, som gør vores produkt unikt. Men vi vil kun foretage indgreb, som er betinget af de arkitektoniske valg, eller betinget af ønsket om øget sikkerhed, fejtolerance mv.

Strategien for styresystemer på Supermax nu og i fremtiden er beskrevet i Ref.3. Valget af alternativ styresystem er beskrevet i SPC/3 rapporten (Ref.1). Den seneste udvikling omkring ACE vedrørende USL's indmeldelse har gjort vores valg mindre kritisk. Der er dog stadigvæk en væsentlig komplicerende parameter, nemlig begrebet **endians**.

Enhver maskine har et 'køn'. Enten er det en *little endian* eller også er det en *big endian*, afhængig af hvordan bytes placeres i forhold til hinanden på busser og i lager. Hardware, styresystem, applikationsprogrammer og data skal have samme 'køn'. PC'er er little endian maskiner, Motorola baserede maskiner er big endian. MIPS baserede maskiner kan være enten det ene eller det andet. De fleste MIPS baserede (inklusiv Supermax) er big endian, DECstations er little endians. Da ARC-specifikationerne foreskriver little endian, er vi nødt til at tage stilling til dette bl.a. i vores styresystemvalg. Problemet er langt alvorligere end det umiddelbart tager sig ud. Der er nemlig ingen lette løsninger på det, og det tvinger os ind

i et brud på bagudkompatibilitet.

Vi bør udvikle hardware således, at det er i overenstemmelse med ARC-specifikationerne. Denne arkitektur er i modsætning til Supermax idag en *little endian* maskine. For at opretholde software og data bagudkompatibilitet ved skiftet af bussen bør hardware laves *Cross-endian*, dvs en måde hvor begge endians supporteres. Dette kan ikke gøres uden en eller anden form for strapning eller fastlæggelse af endians på resettidspunktet.

SMOS, som er big endian, kan derved bringes til at køre på den nye maskine i den første base, med bagudkompatibilitet af software og data.

Mange af ACE deltagerne har i dag (som DDE) big endian maskiner, og står derfor i samme situation som os. Først og fremmest har USL problemet, idet de i dag har SVR4 kørende på MIPS i big endian form. Det er primært DEC, der ikke har problemet. Med USL's deltagelse i ACE skal dette problem løses i SVR4. Hvordan de vil gøre det, vides ikke.

Vores skift til en eventuel binær 'ARC-verden' vil således formentlig indebære et brud med bagudkompatibilitet. Skiftet bør foretages i forbindelse med skiftet til alternativt styresystem. Vi kan så antage at USL har fundet en 'smertefri' måde at håndtere dette skift. Vi kan imellemtiden overveje om der findes en god 'heterogen' løsning på bruddet med fortiden.

Endelig skal man ikke se bort fra den mulighed, at det lykkes AT&T at påvirke sine omgivelser i en sådan grad, at MIPS Risc og big endians alligevel er fremtiden.

3.4. ACE-PC'er

De foreliggende beregninger på de små modellers kostpris viser (ikke uventet), at hvis der er et kostproblem i dag på Supermax i den lave ende, bliver dette ikke bedre på den kommende (se afsnit 5). Når vores focus er på den høje ende af et modulær, UMA/NUMA, 16 CPU system, skal vi have en mulighed for reagere fornuftigt på et prispres i den lave ende. Dette kan klares, hvis vi sikrer at vi kan lægge vores styresystem ind på indkøbte små prisbillige ACE-PC'er, og afvikle (med binær kompatibilitet) alt øvrigt software. Dette skridt kan afvente prisudviklingen i den lave ende, og gennemføres når/hvis det er belejligt. Vi behøver med andre ord ikke placere det i projektforløbet på nuværende tidspunkt.

Der er en række problemstillinger og omkostninger i dette, og den fulde konsekvens af ovenstående skal afklares i forbindelse med valg af alternativt styresystem. Det giver således ingen mening at tage en ACE-PC ind, før vi har foretaget et valg af *endians*.

3.5. Disksystemer

Det intelligente Disk-I/O modul skal kunne tilsluttes dels almindelige diske, dels selvstændige disksystemer.

GogolplexPack, som er omtalt i Ref.1, er en selvstændig diskenhed med indbyggede muligheder for spejling, foldning, sikkerhedskopiering osv.

Vi har ikke beskæftiget os med denne i den mellemliggende periode, men den bør indgå i de kommende undersøgelser vedrørende 'high availability' mm.

4. PRODUKTSTRATEGI

4.1 Udviklingsstrategi

Det er væsentligt i større udviklingsprojekter, som indeholder en lang række teknologiskift, dels at foretage en risikominimering, dels at sikre at eksisterende viden og erfaring er udnyttet og videreudvikles, når gamle produkter kasseres, og nye erstatter dem.

Begge dele kan imødekommes ved en udviklingsstrategi, som indeholder en opdeling af større projekter i en række mindre trin, som fordeler de væsentlige teknologiskift. Disse trin skal sikre, dels at der også under projektforløbet er produkter at sælge, dels at der kan focuseres på at teknologiskiftet faktisk giver de forventede fordele, og at ingen af de eksisterende stærke sider i et produkt sættes over styr.

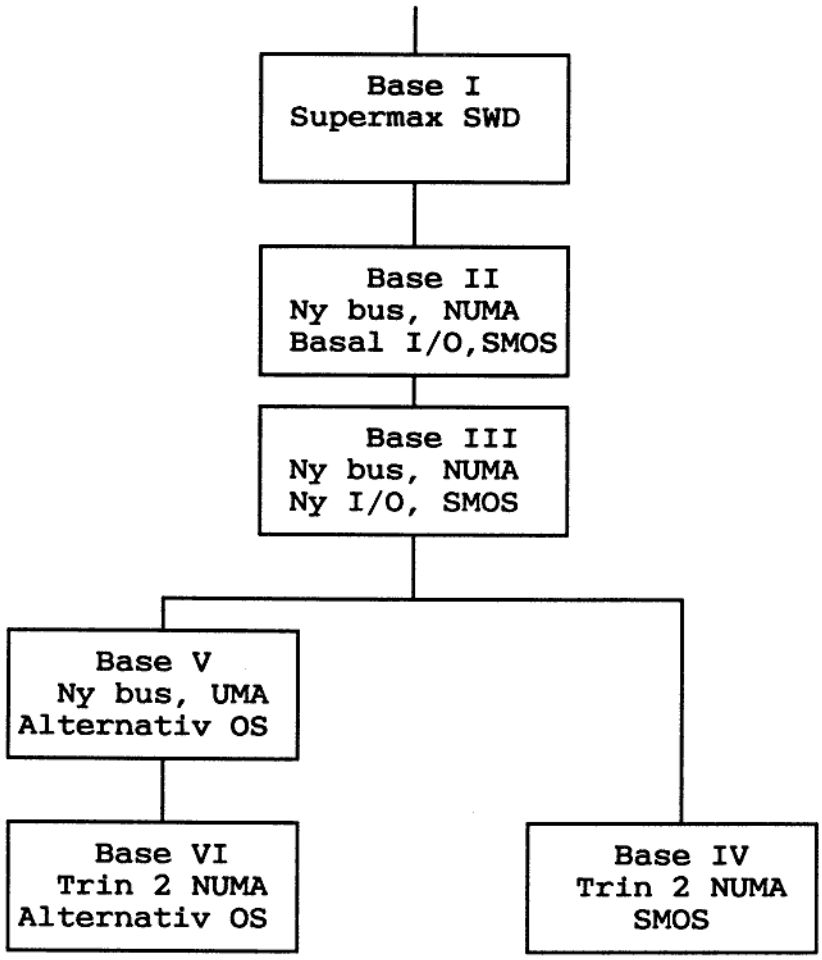
Der er i det følgende brugt udtrykket *base*. Dette udtryk dækker et niveau i udviklingsprocessen, hvor alle definerede moduler (både HW og SW, gamle og nye) er i stand til at fungere i sammenhæng. En base er derfor en milepæl i projektforløbet, som ikke nødvendigvis er det endelige mål. Man kan fra starten eller undervejs vælge om basen skal markedsføres, sælges og leveres (hvilket gør basen til en platform), og man kan undervejs overveje om den valgte vej til de kommende baser til stadighed er den korrekte.

I det følgende afsnit er det beskrevet, hvilke baser dette projekt har, og hvorledes vejen mellem dem kan se ud.

4.2. Supermax Baser

Det er vanskeligt på det nuværende grundlag at definere baserne detaljeret, og det er heller ikke meningen på nuværende tidspunkt. Baserne er valgt udfra en udviklingsmæssig synsvinkel. Platforme kan defineres ud fra disse baser.

Baserne og vejen gennnem dem ser således ud:



Base I: Supermax SPC/3-SWD.

Den første base i udviklingsforløbet er en Supermax RISC platform (med eksisterende bus) med et modificeret SMOS. SMOS modificeres således at kun faciliteter, som også findes i den kommende hardware arkitektur, benyttes. Basen bruges til SW udvikling. Formålet er at udvikle og teste OS inden implementationen på den ny bus og på R4000 materiel. Basen giver ikke mening som en salgbar model.

Base II: Ny bus og NUMA arkitektur.

Næste hardware skridt i videreudvikling af Supermax er en ny bus. Når bussen findes, er et R4000 CPU modul enkel at lave. I/O begrænses (små skridt), hvorfor der startes med Basal I/O modul, som skal bruges i alle modeller. NUMA betyder, at der til hver processor hører et lokalt lager, og at der er et globalt lager.

Samme SMOS som i base I benyttes, men der tilføjes drivere til Basal I/O.

Det store spring er konstruktion af at nyt bussystem. Bussystemet indeholder 2 væsentlige tiltag. Det første tiltag indeholder ændringerne i businterface, elektrisk og mekanisk. Dette giver bussystemet en høj ydelse. Det andet tiltag er cache konsistens, som aflaster bussystemet.

Cache-konsistens er kompliceret. Forskellige begivenheder kan udskyde tidspunktet hvor dette vil fungere. I det følgende antages det, at cache konsistens eksisterer fra denne base. Det skal dog bemærkes at en brugbar Base kan realiseres uden at cache konsistens virker.
Applikationer, der benytter shared memory vil ikke yde godt uden cache konsistens. Kernen skal benytte cache konsistent shared memory og vil yde mindre, hvis denne facilitet ikke er tilstede. Basen vil dog kunne fungere, og indeholder en kraftig processor og I/O ydelse.

Den størst mulige konfiguration ønskes for at kunne teste alle hjørner af hardware i et realistisk SW miljø. På vejen til denne base vil der være en række konfigurationer, der benyttes i udviklingsforløbet. F.eks. vil første konfiguration være et 1 processor system og det næste et 3 processor system.

Basen benyttes til at vurdere, hvormange processorer der kan forbindes til den nye bus. Derimod er I/O begrænset, da kun det basale I/O er udviklet.

Base III: Ny bus og nye I/O controllere.

Denne base består af en Base II konfiguration, med tilføjelse af den fulde I/O båndbredde. Dvs der eksisterer specielle Disk og LAN I/O moduler. Der benyttes SMOS med I/O drivere til I/O modulerne. I denne base er det muligt at konfigurere både store og små maskiner, med et afbalanceret ydelsesforhold mellem CPU'er og I/O.

Base IV: Supermax med Trin 2 NUMA arkitektur.

Denne base består af en Base III konfiguration, hvor CPU'erne kan konfigureres i en NUMA Trin 2 arkitektur. NUMA betyder, at der til hver processor hører et lokalt lager, og at der er et globalt lager. Trin 2 betyder, at det lokale lager kan deles af flere processorer. Der benyttes SMOS med de tilføjelser, der er nødvendigt for at kunne køre i Trin 2 NUMA arkitektur. Samtidig tilføjes I/O bussen ved hjælp af det specielle lagermodul med I/O bro.

Dette er den fuldt udbyggede Supermax med et SMOS styresystem.

Base V: Ny bus, nyt OS med UMA arkitektur.

Denne base består af Base III med R4000 processor- og lagermoduler og I/O moduler i størst mulig konfiguration. Der er anvendt et alternativt styresystem. Det kunne for eksempel være en SVR4/MP, som er udviklet til en UMA arkitektur.

UMA betyder her, at det er en traditionel 'tæt-koblet' multiprocessor med delt lager og cache-konsistens.

Den største mulige konfiguration ønskes for at kunne teste alle hjørner af hardware i et realistisk SW miljø. Basen benyttes til portering af det alternative styresystem.

Base VI: Ny bus, nyt OS med Trin 2 NUMA arkitektur.

Denne base svarer til Base IV, dvs med R4000 processor- og lagermoduler og I/O moduler i den fuldt udbyggede Trin 2 NUMA maskine. Men denne base anvender et alternativt styresystem.

Sammenfattende kan man sige følgende: Base I benyttes til udvikling/test af SW. Base II og III kan blive til en række kommercielle modeller. Erfaringer med disse baser, og ydre forhold omkring alternative styresystemer vil give input til den videre vej mod det endelige mål. Hvis vi på dette tidspunkt vælger at gå over til et alternativt styresystem, skal vejen gå igennem en UMA konfiguration. De to veje ender op med samme materiel. Eventuelt kan man gennemføre et parallelt forløb.

4.3. Mindre relevante baser.

Vi har undersøgt 2 alternative baser, som bygger på den nuværende busteknologi.

Base: Supermax R3000-40.

Basen består af en Supermax RISC platform med SMOS. Der tilføjes RISC- og lagermodul med højere ydelse. Den baseres på R3000 processor med 40 MHz klokfrekvens.

Frekvens forøgelsen skulle ideelt give en ydelsesforbedring på 1,6 gange. Afhængigt af modulvalg og implementering af lagersystem kan der forventes en ydelse på 1,3 til 1,5 gange vores 25 MHz konstruktion. Kost/ydelse forholdet vil muligvis forbedres, men det kræver en væsentlig indsats fra BW/HWD. Små forskelle i implementeringen i forhold til den nuværende konstruktion vil kræve rettelser i SMOS. Software indsatsen vil være lille.

Væsentlige flaskehalse i Supermax forbedres ikke. Tilgangstiden til uncached shared memory er for nogle applikationer (Oracle) meget essentiel og vil ikke blive

forbedret. I/O båndbredden er væsentlig for f.eks. database og server applikationer og vil ikke blive forbedret.

Derfor mener vi ikke, at udbyttet af denne base står mål med investeringerne.

Base: Supermax R4000-50

Basen består af en Supermax RISC platform med SMOS. Der tilføjes RISC- og lagermodul med højere ydelse. Der baseres på R4000 processor.

Single processor systemer, der tillader cache af shared memory vil have en processor ydelse, der er bedre end en Supermax med 2 til 3 R3000 moduler.

Men det er ikke muligt at lave cache-konsistens mellem flere processorer. Det er ikke muligt at lave I/O cache-konsistent. R4000's atomiske operationer (Test and set) vil ikke fungere. Binær portering af applikationer der benytter shared memory og dermed atomiske operationer vil ikke virke.

Nyt lagermodul skal udvikles. Lagermodulet vil ikke kunne genbruge de nuværende RAM submoduler. Kost/ydelse forholdet vil muligvis forbedres, men det kræver en væsentlig indsats fra BW/HWD. R4000 kræver mange rettelser i SMOS.

Ligesom i den før omtalte R3000 base forbedres væsentlige flaskehalse i Supermax ikke. Flaskehalsene vil være relativt større end tidligere, idet I/O ikke skaleres op i samme omfang som processorydelsen.

Vi har ikke medtaget basen, da den ikke er en naturlig vej i produktstrategien. Den giver os ikke nogen væsentlig ydelsesforbedring, den vil blot ramme loftet i systemkapaciteten tidligere.

Hvis der ses bort fra udviklingsressourcer, kan denne base være et skridt frem med lille risiko, men uden ændring af observerede flaskehalse.

5. SKITSE TIL IMPLEMENTATION

5.1. Krav til I/O ydelse

I afsnit 2.1 er omtalt, hvilke miljøer vi forventer Supermax i fremtiden vil indgå i. Med dette som udgangspunkt kan vi opstille følgende krav til I/O ydelse:

Miljø: Oracle database server.

Ydelse ved 16 CPU'er er ca. 600 MIPS (jvf Ref.1). Tallene fra forprojektet er:

- 600 MIPS, 1 TPS ved en 1/2 MIPS, 3 disktilgange pr. TPS

Dette giver 1200 TPS og 3600 disk tilgange pr. sekund. 1 diskblok er 8 kbyte dvs. båndbreddekravet er 3600 * 8 Kbyte = 28.8 Mbyte/sekund.

Den største del af Oracles disktilgange er uden om diskcache, og de disktilgange, som cache's er ofte miss'er.

Miljø: NFS og LM/X server.

SPARCstations og lignende maskiner kan idag mætte et ethernet i NFS miljøer, dvs. de kan belaste med ca. 1 Mbyte/s. Næste generation Workstations, ACE-PC'ere og i586 maskiner vil yde ca. 3 gange den nuværende generation. Der er ingen væsentlige hindringer i softwarelagene for at en ændring i CPU ydelsen ikke skulle give en tilsvarende forøgelse i netværk I/O's data throughput.

Ethernet og Token-ring vil i de kommende år blive suppleret med FDDI netværk. FDDI har et max. data throughput på ca. 10 Mbyte/s. SPC/3 vil kunne understøtte mere end 1 FDDI net. Data til FDDI nettet er typisk data fra disk-cache eller disk. Hvis data kommer fra disksystemet skal data først transporteres fra diskmodul til globalt lager og dernæst til FDDI-modul, hvilket kræver en busbåndbredde, der er mindst 2 gange netværkets datathroughput.

CPU ydelse er vokset meget over de seneste år. Diske og LAN I/O ydelse er ikke vokset i samme omfang som CPU ydelsen. Der er dog tre forhold, der vil påvirke kravene til I/O systemer over de næste år. Det første forhold er fremkomsten af disk-systemer, hvor diskene udnyttes på en innovativ måde, hvilket øger kravene til båndbredde. Det andet forhold er forventningen til udbredelsen af FDDI, hvor netværksbåndbredden 10-dobbles iforhold til ethernet/tokenring. Det tredie forhold er brugen af SPC/3 i ovennævnte miljøer, hvor der vil stilles store krav til responsetid, throughput og kapacitet. Her er det udviklingen omkring applikationer og kravene til grafisk datarepræsentation, der vil stille krav til I/O systemet.

Konklusionen på disse overvejelser om kravet til I/O ydelse er den følgende:

De fleste miljøer vil være en blanding af de tre beskrevne. Samtidigt stiller ønsket om høj systemtilgængelighed krav om sikkerhedskopiering under drift. I/O systemet skal konstrueres således at ovenstående datamængder kan håndteres med et passende fremtidigt råderum.

I/O systemets mål er at kunne levere effektivt 50 Mbyte/s I/O data.

5.2. I/O Valg

I det følgende gennemgås argumentationen for nøglevalgene omkring I/O, der er nævnt i Afsnit 3.

Der er to faktorer, der taler for at adskille I/O moduler fra systembussen:

1. Et ønske om at kunne skalere op til endog meget store systemer.

2. En forventning til at I/O systemet ikke vil gennemgå samme teknologiskift som CPU'erne og dermed have en lang levetid.

Forudsætningerne om skalerbarhed og modularitet betyder at I/O systemet er baseret på et bussystem. Adskillelsen opnås ved at have dels en systembus og dels en I/O bus.

Kravet til ydelse betyder, at der kun er få valgmuligheder med hensyn til I/O bus. Samtidige ønsker om at få mulighed for at tilslutte 'standard' I/O moduler (f.eks EISA) kan honoreres på forskellig vis, og dette er ikke nødvendigvis o et krav til I/O bussen.

I/O busser:

Den eneste "standard" bus der kan honore ydelseskravene er Futurebus+ profile B. Denne bus er et resultat af et omfattende fælles specifikationsarbejde af en lang række leverandører. Men FB+ produkter findes endnu ikke. Det forventes, at DEC som en af de første vil annoncere produkter, der benytter FB+ i løbet af 2. kvartal 1992. Det er endnu uvist i hvilket omfang, og hvornår der vil forefindes et passende udbud af standard produkter. FB+ vil gennemgå en indlæringsperiode både teknisk og kommercielt. FB+ implementeringen er omfattende og teknisk avanceret, og dette vil betyde, at produkterne i en periode vil være dyre. En implementering af FB+ i SPC/3 vil være muligt, men der vil ikke indenfor en forudsigelig fremtid være et passende udbud af standard I/O moduler.

Alternativt til FB+ kan der implementeres en DDE specifik I/O bus. En DDE specifik I/O bus vil kunne honorere ydelseskravene, men vil tilgengæld ikke opfylde ønsket om tilslutning af standard I/O moduler.

Vi har valgt en DDE-specifik I/O bus, og vi har valgt at

denne bus har samme specifikationer som systembussen.

Dette giver en række fordele. I/O moduler kan placeres i samme bus som CPU-moduler og lagermoduler, hvilket giver maskinen og projektet en stor fleksibilitet. Ulemperne er primært at en 64 bit bred bus til I/O er bedre ydende end absolut nødvendigt og har dermed en højere kost end nødvendigt. Bemærk dog, at det er en fordel for de små maskiner. De små maskiner kan nøjes med et backplane, et CPU-modul, et lagermodul og et I/O modul.

Tilslutning af standard I/O moduler omtales senere.

I/O moduler:

Primær I/O tilsluttes ved hjælp af intelligente **I/O moduler**. Det er Disk I/O og LAN I/O, som indeholder en lokal processor. Den lokale processor aflaster CPU-modulerne. De nederste lag af I/O driverne og dermed det programmel, der håndterer et stort antal interrupts, eksekveres lokalt i I/O modulerne. DMA kanaler placeres på de enkelte moduler, for at opnå en ideel skalering når flere moduler adderes. Der er mulighed for, at processorer og I/O moduler kan sende interrupt til hinanden.

Derudover skal CPU-modulerne og I/O modulerne kommunikere gennem et *delt lager*.

Dette delte lager kan placeres i det globale lager eller i det enkelte I/O modul. Kravene til I/O modulernes funktionalitet er vidt forskellig i de to tilfælde. I det første tilfælde er I/O modulets lager *single-port* og i det andet er det *dual-port*.

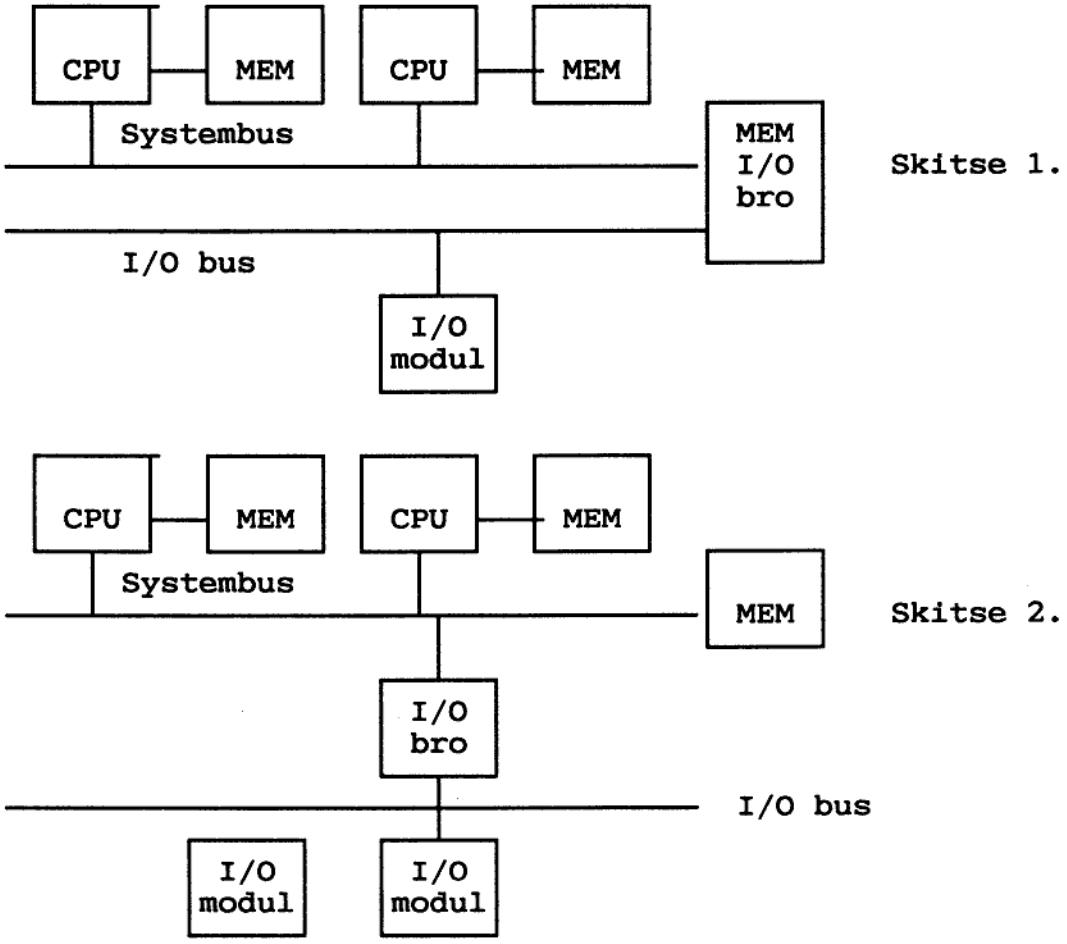
Single port moduler er simplere at implementere i hardware. I/O moduler vil altid transportere data til/fra det globale lager, og delte data er kun placeret her. Kommunikationen mellem applikationsprocessor og I/O modul skal benytte interrupt, idet det ikke er muligt at overvåge I/O modulet fra CPU-modulerne ved at polle strukturer i det globale lager.

Dual-port moduler er mere komplekse at implementere i hardware. Delte data kan placeres hvor man ønsker det. Data kan flyttes af CPU-modulerne fra/til I/O modulerne, men med en begrænset ydelse idet data ikke flyttes som cache-lines. Der kan benyttes pollede strukturer og/eller interrupt.

For at forenkle har vi valgt single port moduler.

Systembussen og I/O bussen kan som nævnt være to seperate busser. Disse to busser kan som vist på nedenstående to skitser forbindes på to måder. Forskellige årsager, bl.a. anvendelsen af disk cache gør, at alle I/O data altid er placeret i det globale lager før/efter en I/O operation. I skitse 1 er I/O bussen derfor forbundet direkte til det

globale lager. I skitse 2 er I/O bussen forbundet til systembussen.



Skitse 1 fordele: Der benyttes et færre antal slots i systembussen. Hvis busbåndbredden er en flaskehals mere end lagerbåndbredden yder dette system mere end skitse 2. Hvis I/O ikke benytter den cache-konsistente systembus bliver kravene til denne mindre.

Skitse 1 ulemper: I NUMA konfigurerede maskiner er der ingen mulighed for at læse/skrive I/O til det lokale lager. Beslutningen om at I/O læsning/skrivning er til det globale lager er dybt begravet i hardware arkitekturen. Det er ikke muligt at lave cache-konsistent I/O. Der skal udvikles et specielt lagermodul til store maskiner.

Skitse 2 fordele: Der er ingen valg, der hindrer I/O adgang direkte til det lokale lager, hvis dette viser sig at være relevant. Der er mulighed for at I/O er cache-konsistent. Det kan have betydning for kommando/status I/O data, der kan cache's og dermed kun belaster bussystemet med cache-line transaktioner istedet for et stort antal single transaktioner. Hvis lageret og ikke systembussen er flaskehals, er skitse 2 ligeså god som skitse 1.

Skitse 2 ulemper: Der skal udvikles en bridge. Der benyttes båndbredde på systembussen.

Vi har valgt skitse 1.

Basalt I/O:

I alle systemer indgår et **Basalt I/O kort**. På små maskiner er dette kort eneste I/O. På større maskiner har det primært til formål at huse forskellige fælles centrale ressourcer. Disse fælles ressourcer er f.eks. non-volatile memory, calender clock, og andre ARC-specificerede faciliteter. Det basale I/O indeholder derudover: processor, lager, businterface, boot-proms og I/O funktioner. I/O funktionerne understøtter SCSI, LAN og kommunikationsfunktionerne ISDN og HDLC. Kortet tilsluttes altid til systembussen.

Mulighed for standard I/O kort:

Tilkobling af standard I/O bus er nødvendig for at kunne tilslutte eksotiske I/O devicer (scanner, fax, talegenerering, krystalkugle...). Med ACEs styresystemer og SVR4 er I/O driver interface til styresystemet veldefineret, og devicedriveren kan derfor forventes leveret sammen med hardware. Disse kort/driver systemer må forventes at komme til ACE-PC'er (ARC). Hvis de laves til Compaq ACE-PC bør vi have en EISA bus. Hvis de laves til DEC bør vi have en TurboChannel(TC). Det er de to busser ARC pt. specificerer

EISA er af mekaniske grunde ubehagelig at implementere i SPC/3. EISA kan benytte både ISA og EISA moduler. ISA modulerne har kun et 24 bit adresserum hvilket vanskeliggør implementeringen yderligere. TC er en "sandwich" type I/O modul og kunne være en oplagt kandidat til det basale I/O modul. TC svaghed er måske mangel på udbredelse.

Tilkobling af TC og EISA kræver begge, at de kan adresseres fra en processor der eksekvere en I/O driver. Processoren kan være en I/O processor eller et CPU-modul.

Det er sandsynligt, at implementationen kræver at driveren eksekveres af CPU-modulet. Denne må have direkte adgang til I/O modulerne. En mulighed er at placere standard I/O broen som et modul på systembussen uden lokal processor.

Beslutning om hvilken standard bus, der skal anvendes, er indtil videre åben. Beslutningen vil påvirke det mekaniske design.

5.3. Backplane-teknologi og businterface.

Et shared-memory multiprocessor system med mere end 4 processorer kræver brug af backplane-teknologi. Der kræves to væsentlige egenskaber, der begge er svære, men mulige at

honorere. Den første egenskab er cache-konsistens (dvs. at holde alle cache-lagre konsistente), som gør systemet logisk komplekst. Den anden egenskab er, at gør backplane kommunikationen hurtig og pålidelig. Her er det især de elektriske egenskaber, der er svære at honorere.

Der er 4 mulige veje:

1. Futurebus+ med profile A eller F conformance.
2. Baseret på silicon fra en R4000 leverandør eller lignede silicon producent.
3. DDE specifikt synkront system.
4. DDE specifikt synkront system med brug af Packet Data Fifos (PDF).

Ad.1. Er forkastet pga. ydelse, tilgængeligheden af silicium og muligheden for effektivt at interface R4000 til FB+.

Ad.2. Undersøges løbende. Pt. findes kun eet projekt. Projektet drives af Prime og LSI-Logic i fællesskab og består af et chip-set. Komponenterne muliggør et system som meget ligner SUNs MP system. 1 til 4 processorer i systemet forbundet vha. en board level bus. Processorer og cache er placeret på 1 til 2 plug-in moduler. Moduler med fremtidige udgaver af R4000 kan benyttes.

Ad.3/4. Er realistiske løsningsmuligheder og vil blive nærmere behandlet i det følgende.

DDE specifikt system:

Den omfattende indsats omkring FB+ har tilvejebragt en del meget brugbar teknologi som kan benyttes af DDE. Fordele ved at benytte FB+ teknologi inkluderer dokumenterede standarder, tilgængelig silicon og andre dele som vil blive produceret i stort styktal samt en garanti for nye produkter i fremtiden.

Elektrisk interface:

Baseres på incident wave switching som udføres vha. FB+ BTL teknologi. Al kommunikation er synkron fra register til register. BTL muliggør hurtig kommunikation over backplane.

Mekanik:

Al mekanik er baseret på FB+'s metriske standard. PCB moduler, frontpanel, EMC tætning, miljø-krav etc. kan udføres i henhold til FB+ standarden. Benyttelse af disse standarder betyder at en masse forhold er veldokumenterede, og at tingene hænger sammen.

Busprotokol og funktionalitet:

Bus protokollen baseres på de funktioner R4000 tilbyder, det arbejde der er skitseret i R4050 projektet og de krav og

input, der kommer fra anden side. Input fås fra ARC-MP specifikationen, Prime's Mustang projekt og SUN international's MBus specifikation. Cache konsistens er implementeret vha. en snoopy protokol, og det undersøges, om det er ønskeligt og muligt at benytte duplicate tags. Selve protokollen og transaktionerne er endnu ikke defineret.

Implementering af businterface:

Businterface er interfacet mellem den cache-konsistente system bus, processor-moduler og lager-moduler. Businterfacet kaldes også en External Agent. Protokollen er meget kompleks, og antallet af nødvendige registere/adresseveje er så stort, at det er nødvendigt at benytte ASICs for at kunne implementere en external agent.

Busprotokollen kan implementeres med en mere eller mindre avanceret protokol. Den mest effektive protokol vil have support for split transaktioner og posted writes. Den detaljerede implementation vil være et "trade off" mellem de ideelle egenskaber og de reelle muligheder.

Interfacet mod processoren udgøres af et synkront interface. Det er målet at businterface skal kunne benyttes af flere generationer R4000 processorer. Nye generationer vil dog kræve redesign af CPU-modul. Næste generation processorer vil være baseret på 3.3 Volt forsyningspænding.

Interfacet mod den cache konsistente bus udgøres af en synkron bus på 25 eller 33 MHz. Datavejen udgøres evt. af FB+ packet data fifos, PDF. PDF er komponenter, der kan flytte data i blokke med stor hastighed. PDF er ikke velegnet hvis der ønskes et lille antal processorer på en bus. PDF giver båndbredde, men øger tilgangstiden.

Den overordnede arkitektur er baseret på 2 businterface på hvert modul. Ønsket om 2 businterfaces stiller store krav til implementeringen. Der tilføjes en del kompleksitet, og der adderes lidt til tilgangstiden til lageret.

Med den omtalte teknologi kan følgende båndbredder opnås. Båndbredden er med en avanceret busprotokol og derfor meget optimistisk.

Anslået båndbredde	uden PDF	50 MHz PDF	75 MHz PDF
1 interface 64 bit:	224 MB/s	284 MB/s	387 MB/s
2 interfaces 64 bit:	358 MB/s	454 MB/s	620 MB/s
1 interface 128 bit:	387 MB/s	426 MB/s	533 MB/s

Det er ikke muligt at have 2 128 bit interfaces på et modul. 128 bit interface viser ydelsen for en alternativ implementation med kun 1 interface. 2 interface værdierne er for en UMA arkitektur hvor de to 64 bit bredde busser logisk benyttes som en bus.

Implementeringen af businterfacet vil være et trade-off. PDF konstruktionen giver båndbredde men øger tilgangstid til lageret. PDF er derfor velegnet til UMA arkitektur men mindre egnet til NUMA arkitektur. Det sandsynlige valg er at lave en synkron konstruktion uden PDF.

Det næste problem er så hvad denne båndbredde kan benyttes til. Det er ikke umiddelbart muligt at lave eksakte målinger, simuleringer eller undersøgelser. Kravene til båndbredde vil være stærkt applikationsafhængige og med en række nye R4000 processorer i udsigt er det ønskeligt med så stor en båndbredde som overhovedet muligt. En styrke ved NUMA arkitekturen er at antallet af processorer tilsluttet en lokal bus er lille og at systemets samlede busbåndbredde er stor.

Backplane:

Det maksimale antal slots i maskinens backplane er ved et synkront 33 MHz system anslået til ca. 24. Et sådant backplane vil kræver montering af moduler på begge sider af backplane. Antallet af slots er begrænset af hvor store PCB moduler, der kan produceres, impedance forhold og løbetid.

Hvis der benyttes PDF's ved 75 MHz for at øge busbåndbredden anslås det, at det maksimale antal slots er ca. 16.

Der skal benyttes et slot til et utility modul, der indeholder nødvendig klok distribution og andre centrale faciliteter.

Det er nødvendigt at SPICE simulere (analog elektrisk simulering) eller at eksperimentere for at eftervise backplanets funktion. Det kan være nødvendigt at have "filler panels" i tomme slots, dvs. at tomme slots elektrisk belaster bussen således at bussen er en hel homogen transmissionslinie.

Grundlaget for backplane konstruktion er FB+ specifikationer. De foreløbige FB+ backplanes er konstrueret som 16 layer PCBs. Der er 3 signallag, resten benyttes til low impedance VCC/GND + lidt andet. Konnektorerne benytter pressfit teknologi, og termineringerne er udført med SMT komponenter. Backplanet er tykt pga. impedans forhold. Tykkelsen giver produktionsproblemer pga. den relativ lille hul diameter der skal benyttes til konnektorer med meget høj densitet.

Backplane med FB+ egenskaber er udviklet af enkelte producenter. Vi vil få demonstreret nogle produkter i november. Backplane med de krav der stilles, vil kunne implementeres. Det er muligt at hente information/teknologi og samarbejde med BICC-VERO, som allerede har en pilot produktion af FB+ backplanes. En stor del af BICC-VEROs virksomhed beskæftiger sig med udvikling og produktion af backplanes, f.eks. backplane til ICL DRS6000 SPARC system.

5.4. Elektronik modulerne

En del af modulerne indeholder et businterface. I det følgende beregnes det benyttede PCB areal til et 64 bit businterface med PDFs.

De enkelte komponenter:

ASIC	:	45 x 45 sqmm	=	2210 sqmm.**
Register	:	11 x 16	=	240
PDF	:	24 x 18	=	520
BTL	:	13 x 13	=	225
R4000 socket	:	65 x 75	=	5160
SRAMS	:	19 x 12	=	300

** 2 mm border. Dvs. 4 mm mellem IC'er

1 64 bit bus interface:

1 ASIC	:	2210
8 registers	:	1920
8 PDFs	:	4160
10 BTL	:	2250
ialt		10540 sqmm.

Processor modul:

Modulet indeholder R4000 processor, 2x64 bit businterfaces, 1 MB sekundær cache og diverse logik til klok, reset etc.

1 R4000	:	5160	
44 SRAMS	:	13200	
2 64 bit if.	:	21080	
diverse	:	7200	(30 registers)
ialt		46640 sqmm	

FB+ board areal: 265 x 287.5 = 76187 sqmm
Kortet kan realiseres.

Lager modul:

Modulet indeholder 1 eller 2x64 bit businterfaces, EDAC, lagerstyring og DRAM array. Lageret indeholder 128 Mbyte lager ved brug af 4 Mbit teknologi. Det er sandsynligt at 16 Mbit teknologi vil være tilgængelig ved SPC/3s salgstidspunktet. Med 16 Mbit teknologi kan der være 512 Mbyte pr. modul.

Businterfaces	:	1/4 board areal
EDAC+styring	:	1/4 -"-
DRAMS	:	1/2 -"-

DRAM er i zig-zag pakning. Alternativt kan der benyttes

SIMM.

Clock/utility modul:

Der skal distribueres klok signaler rundt til alle moduler. Tilgængelige controlled skew clock drivers skal benyttes. Klok signalernes længde i backplane skal tilpasses så alle er lige lange og belastet. Power-up reset, elektronik for betjeningspanel og lignende er funktioner der håndteres af utility modulet. Arbitrering for adgang til busserne forsøges distribueret ud på de enkelte businterfaces. Den store flexibilitet i arkitekturen hindrer een simpel central arbitrering.

5.5. Portning af SMOS til SPC/3

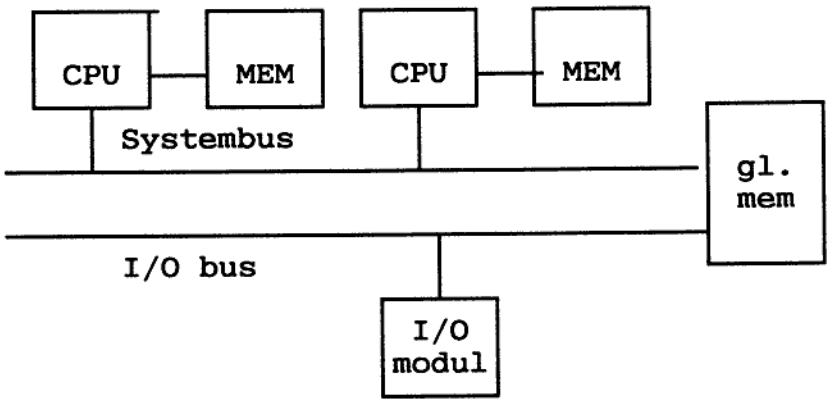
I det følgende gennemgås nogle af de programmel opgaver, der følger af SPC/3's specielle arkitektur. Opgaverne beskrives i hver base for sig.

Hvert opgave er klassificeret i 3 klasser:

SKAL LAVES	Kode der skal skrives.
VIGTIGT	Maskinen virker, hvis dette ikke er lavet, men i en eller anden forstand uacceptabelt dårligt.
GODT	Har væsentlig positiv betydning for ydelse/egen- skaber.

Base I-III:

SMOS portes først til denne `extreme NUMA maskine':



Denne maskine adskiller sig fra en Supermax ved følgende egenskaber:

- Den ene CPU kan ikke se den andens lokale lager.
- I/O kortene kan ikke se lokal lagrene.
- Det er ikke muligt at `tasse' I/O enheders lager.
- Fællesbussen er væsentligt højere ydende end på Supermax.
- I/o kortene er væsentligt anderledes end på en Supermax.

Dette giver anledning til en række ændringer i SMOS: (her indgår en del lokal slang)

* Globalt tilgængelige items flyttes til fælles lager. (SKAL LAVES 2 måneder)

* Globalt tilgængelige data i pl'er flyttes til fælles lager (SKAL LAVES 1 måned)

* I/O til bruger data foregår via fællesbussen. (fit/freebuf) (SKAL LAVES 2 måneder)

* Evt. skal terminal/net I/O foregå via AT&T's MP streams. (GODT 2 måned)

* Al disk I/O skal foregå via disk cache i fælleslagret. (SKAL LAVES 2 måneder)

. Ny interrupt håndtering. (SKAL LAVES 2 måneder)

. Nye I/O drivere skal skrives til nye I/O kort. (SKAL LAVES 13 måneder)

. Behandling af R4000 nye tlb og lagersystem skal skrives. (SKAL LAVES 2 måneder)

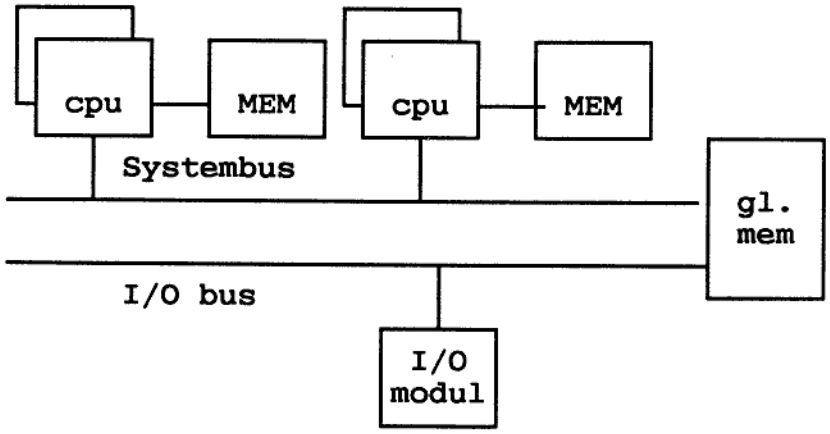
De med * mærkede rettelser kan testes på vores nuværende Supermax hardware (Base I). (Om end det ikke vil bevirke, at den kører hurtigere.) Tiderne for I/O systemet er de mest usikre, men er sat til det samme som for SVR4/MP til Base V.

Base III omfatter tilslutning af primære I/O moduler. (6 måneder.

Base IV:

Ændringer for at understøtte ægte Trin 2 maskiner:

Når UMA egenskaber ønskes vil arkitekturen se ud som følger:



Her er opnået en række fordele:

- Bedre afbalancering af processor kraft.
- Bedre afbalancering af lagerforbrug.
- Mindre maskiner bliver billigere.

Vores ønske er her at opnå en balanceret udvikling ved at kombinere UMA arkitekturen med den globale NUMA arkitektur.

Indførsel af UMA egenskaber kræver rettelser i:

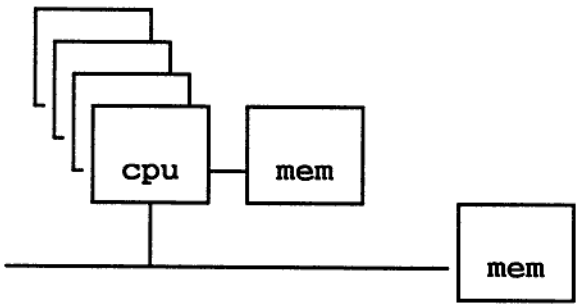
- Lagerallokering (SKAL LAVES 2 måneder)
- Scedulering (SKAL LAVES 2 måneder)
- Reservation omkring tidligere `private' data:

  Item.
  Text beskrivere.
  Regions beskrivere.
  Lager beskrivere.
  (SKAL LAVES 2 måneder)

- Rettelser omkring interrupt rutiner, hvis det ikke er fast hvem, der får interrupt. (SKAL LAVES 2 måneder)

5.6. Portning af SVR4/MP til SPC/3

Base V - UMA maskinen:



(Estimaterne findes i Ref.1).
Egenskaber der kan give årsag til ændringer i standard SVR4-ES/MP kernen og umiddelbare omgivelser:

1: To busser og ikke én.
2: Egne i/o kort.
3: Mips CPU ikke 386.
4: Evt ikke usynlig bro til standard bus.
5: Man kan frygte, at vi har et andet afviklings-miljø end det SVR4-ES/MP kernen er tunet til. (evt mange flere processer, der hver især er mindre.)

Følger af Base V maskinen egenskaber:

ad.1: Dette skulle ikke direkte være synligt for programmellet, men der vil være nogle indirekte følger:

a. Vi bør sikre os, at allokerede lagerblokke fordeler sig pænt på de 2 busser. (VIGTIGT)

b. Da vi forhåbentlig har mulighed for flere CPU'er, bør vi sikre os tilstrækkelig finkornet låsning. (GODT)

ad.2: Vi må skrive eget interface til egne kort. (SKAL LAVES)

ad.3: Selv om kerne m.m. er skrevet i c, bevirker skift af CPU en del arbejde:

a. MMU systemet skal skrives om. (SKAL LAVES)

b. Cache opbygget anderledes:

I: Ændre strategi for cache-konsistens. (SKAL LAVES)

II: Evt ændret strategi for lagerside allokering. (VIGTIGT)

c. Kernen skal gennemgås for steder hvor man udnytter at visse operationer på 386 er atomiske, operationer der ikke er atomare på en RISC prosessor. (f.eks: i++). (SKAL LAVES)

ad.4: Standard drivere kan ikke umiddelbart benyttes. (SKAL LAVES)

ad.5: På fler-CPU maskiner med mange processer i forhold til antal CPU'er er det centralt, at processerne så vidt muligt bliver på samme CPU, for at cachen kan holdes "fresh and hot". Virkningen af at processer hopper meget mellem CPU'er kan være katastrofal på ydelsen. Det er ikke sandsynligt at AT&T(USL) har gjort noget for at undgå dette. (VIGTIGT)

Base VI: NUMA maskinen:
Egenskaber der kan give årsag til ændringer i standard SVR4-ESMP kernen og umiddelbare omgivelser:

1: Ikke alt lager kan ses af alle CPU'er.
2: Mere buskapacitet til "lokalt" lager end til "globalt".
3: I/o sidder kun på den ene bus. (den globale)

ad 1 og 2: Disse 2 punkter kræver bevidsthed om hvilket lager, der anbringes hvor.

Den letteste måde er nok at gå frem i 3 skridt:

Skridt 1:

Få et kørende system på alle CPU'er. Dette system vil ikke udnytte HW godt p.g.a. skæv bus belastning. Systemet har ikke interesse i sig selv.

a. Stack, kode, data allokeres lokalt, alt andet globalt. Specielt alle OS data allokeres globalt. (SKAL LAVES)

b. Processer mærkes med hvilken række deres stak, kode og data ligger i. Når en processor skal vælge næste proces vælger den kun blandt processer i rigtig række. (SKAL LAVES)

c. Der rettes i swap, bl.a. aktiveres der en swap-process (hvad den nu hedder) på hver række. (en pr. lokalt lager). (SKAL LAVES)

d. Det sikres, at disk I/O altid foregår til/fra den globale disk cache.

e. Proc pseudo filsystemet bør rettes, måske bør dette flyttes til skridt 2.

Skridt 2:

Flyt nogle OS-data til lokalt lager. Målet her er at flytte så meget så muligt til lokalt lager for at aflaste global bussen, dog ikke så meget at vi får et væsentlig overhead. Her benytter vi vores erfaringer fra SMOS.

f: Adder strukturer, der beskriver opdelingen på lokalt og globalt lager. (SKAL LAVES)

g: Flyt den proces specifikke "os stack" til lokalt lager. (VIGTIGT)

h: Flyt mmu-træet til lokalt lager. (VIGTIGT)

i: Lav icc mekanisme bl.a. til at accesse lager på andre CPU `strenge'. (SKAL LAVES)

j: Check at streams og andet ikke refererer brugerdata

(primært brugerens os-stack) fra fremmede CPU'er. (SKAL LAVES)

Skridt 3:

Flyt SHM til lokalt lager. Formålet med dette skridt er at lægge SHM på lokalt lager indtil det accesses fra andre `rækker'. Dette skridt skal kun tages, hvis der opstår behov, f.eks kunne man tænke sig, at fremtidige oracle versioner kunne drage fordel af det.

k: Lager allokering til delt lager skal ske lokalt. (SKAL LAVES)

l: Page fault routinen skal flytte delt lager til fællesbussen. (SKAL LAVES)

m: En tilbageflytnings proces skal flytte delt lager tilbage ved festlige lejligheder. (SKAL LAVES)

6. KOSTPRISESTIMATER.

I det følgende betegner en SPC/3 en Base II Supermax. Dvs. en R4000 baseret Supermax med ny bus. Alle prisestimater er så vidt det er muligt baseret på 100 stk. om året i 1992.

6.1. Businterface.

Modulerne til SPC/3 kan forsynes med to businterfaces. Det følgende er en kort beskrivelse af en businterface samt et estimat af prisen.

Bussen er 64 bit bred og anvendes både til overførsel af adresse og data.

Den intelligente del af businterfacen er samlet i et gate array. Der anvendes to gate arrays på alle kort, som er tilsluttet SPC/3 busserne. Prisen på et gate array er ved 1000 stk. mellem 700 kr. og 1.190 kr., hvortil skal lægges prisen for opstart af gate array produktionen.

Opstart af gate array produktionen skønnes at koste 1.000.000 kr. Dette er et gæt. Hvis DDE's leverancer starter et år efter opstarten er betalt og strækker sig over de følgende tre år, hvor der anvendes 2000 gate arrays/år, vil opstartomkostningen beløbe sig til 235 kr./gate array ved en rente på 12% p.a. Prisen på et gate array sættes i det følgende til 1200 kr. inklusiv opstart af produktionen.

Øvrige businterfacekomponenter sættes til 2.400,-.

Pris for et businterface:

1 stk. gate array a	1.200	1.200
Øvrige buskomponenter		2.400
8 stk. BTL transceiver a	50	400
Total		kr. 4.000

6.2. CPU-kort (1 Mb cache, 2 businterfaces):

De dyre komponenter på CPU-kortet er, foruden R4000 og de to businterfacer, cache lageret. Cache-lageret består af 44 stk. statisk RAM 64K4 10 ns, ialt 1Mb data plus diverse. Prisen for 64K4 er 210 kr. ved 10.000 i 1992.

Prisen for et printkort sættes til kr. 2200, hvilket er det samme som vi giver for 4501, det nuværende RISC CPU-kort.

Prisen for diverse komponenter er sat til kr. 3000. Diverse omfatter alle de komponenter, hvis funktion man ikke umiddelbart kan gennemskue. Dette tal er skønnet ved at undersøge vores nuværende konstruktioner.

Pris for et CPU-kort:

1 stk. R4000	7.000	7.000
44 stk. 64K4 statisk RAM	210	9.240
2 stk. businterface	4.000	8.000
1 stk. printkort	2.200	2.200
diverse		3.000
Total		kr. 29.440

Prisen på CPU kortet domineres af prisen på kun tre komponenter: R4000, 64K4 RAM og businterfacerne. Priserne på disse komponenter udgør ca. 24.000,- ud af en samlet pris på 29.000,-.

"Low" cost CPU (256 kb cache 1 businterface):

1 stk. R4000	7.000	7.000
44 stk. 16K4 statisk RAM	60	2.640
1 stk. businterface	4.000	4.000
1 stk. printkort	2.200	2.200
diverse		3.000
Total		kr. 18.840

6.3. Lagerkort.

Lagerkortet er tovejs-interleavet med en ordlængde på 64 bit plus ECC. Lageret er opbygget ved hjælp af dynamisk RAM 4M i zig-zag pakninger. Minimum lagerstørrelse: 64 Mb. Maksimum lagerstørrelse pr. kort: 128 Mb. Lagerkortet er forsynet med et eller to businterfaces.

128 Mb RAM med 2 bus interfaces:

288 stk. HM 514100AZ-8	115	33.120
1 stk. EDAC	1.300	1.300
2 stk. businterface	4.000	8.000
1 stk. printkort	2.200	2.200
diverse		3.000
		kr. 47.620

Prisen på RAM-lager svarer til vores nuværende pris. Businterfacerne udgør igen en betydelig del af prisen.

"low" cost lager: 64 Mb med et businterface:

144 stk. HM 514100AZ-8	115	16.560
1 stk. EDAC	1.300	1.300
1 stk. businterface	4.000	4.000
1 stk. printkort	2.200	2.200
diverse		3.000
		kr. 27.060

6.4. Basalt I/O-kort.

Kortet indeholder al I/O til en lille maskine. Kortet fremstilles kun med en interface til bussen. Prisen for kortet

skønnes at være det samme som kostprisen på en MIOC/DIOC III plus et businterface:

Basalt I/O	8.000
1. stk businterface	4.000
	kr. 12.000

Bemærkninger:

Kostprisen for en SPC/3 med en processor vil være højere end for en tilsvarende en-CPU-maskine, primært svarende til prisen på businterfacen og backplane system. Dette er prisen for skalerbarheden.

6.5. Kostpris for lille SPC/3.

SPC/3 er konstrueret med store maskiner for øje. Det er derfor særligt interessant at sammenligne den mindste SPC/3 med den mindste Supermax.

Supermax:

1. stk. 4501 RISC CPU	10.318
1. stk. 4400 Memory Module	1.543
2. stk. 3300 4 Mb RAM	2.794
1. stk. 4001 DIOC III	8.176
1. stk. 4600 MIOC med I/O	9.192
Subtotal	kr. 32.023
1.2 Mb diskette	497
380 Mb winchester	6.608
525 Mb streamer	3.248
Kabinet SM4	14.000
Supermax total	kr. 55.376
Cost/Performance (20 Mips)	2.669 kr./MIPS

SPC/3:

1. stk. low cost CPU	18.840
1. stk. low cost 64 Mb RAM	27.060
1. stk. basalt I/O	12.000
Subtotal	kr. 57.900
1.2 Mb diskette	497
1.5 Gb winchester	8.000
Streamer	3.248
Kabinet SM4	14.000
SPC/3 total	kr. 83.645

Cost/Performance (40 Mips) 2.091 kr./MIPS

Bemærk at hovedlageret for SPC/3 er 8 gange så stort som hovedlageret for Supermax, og disk kapaciteten 4 gange så stor, samt at ydelsen også er dobbelt så stor.

Det er væsentligt at være opmærksom på følgende:

- Komponentpriserne er estimeret for 100 stk i 1992. Prisen dækker derfor begyndelsen af komponenternes priskurve, og de kan derfor forventes at falde væsentligt.

- En væsentlig del af kostprisforskellen mellem de to små maskiner udgøres af lageret. Anvendelse af 64 Mb lager vil formentlig også på introduktionstidspunktet være mere end rigeligt. Man kan derfor overveje, om der skal laves et specielt lille billigt lager til de små maskiner.

7. PRODUKTION, TEST OG SERVICE.

Den måde en Supermax testes, produceres og serviceres påvirkes ikke væsentligt af det skitserede generationsskift. Dog skal man være opmærksom på følgende:

Produktion: Det er en forudsætning ved konstruktionen af maskinen, at der kan anvendes en blanding af overflademonterede komponenter og traditionelt monterede komponenter på printene. Overflademonterede komponenter kan placeres på begge sider af printet.

. DDE's produktion kan ikke producere denne type printkort idag. Der skal derfor investereres i produktionsudstyr og uddannelse af produktionen. Alternativt kan produktionen foretages af underleverandører.

. Bussen skiftes ud og alle testopstillinger skal derfor naturligvis skiftes. Kortstørrelsen bliver en smule mindre end i dag.

Test: Testen i produktionen sker principielt på samme måde som idag. Det skal dog overvejes om incircuit-testeren har den nødvendige kapacitet og de nødvendige egenskaber.

Service: Service foretages som idag. Fejlen lokaliseres til modulniveau ved hjælp af testprogrammer og error log. Det fejlbehæftede modul udskiftes og repareres i produktionen. Udskiftning af komponenter i marken er generelt umulig.

. Den nye maskine vil være opbygget på samme måde som Supermax i dag, modulært og skalerbart. Forskellene vil ligge i, at der kan bygges større maskiner med større antal CPU'er koblet sammen både UMA og NUMA. Der kan derfor være behov for flere testværktøjer på integrationsniveau.

8. ØKONOMI

Udviklingen af hardware medfører omkostninger til prototyper, gate arrays, testudstyr og kabinetter. Disse udgifter er estimeret til 3.5 mill. kr.

Udgiften er estimeret på følgende måde:

Proto-typer:

Hardware-udvikling.
- 2 maskiner med 1 CPU 1 lagerkort og Basalt I/O
- 1 maskine med 2 CPU'er 2 lagerkort og Basalt I/O

Operativsystem-udvikling.
- 1 maskine med 1 CPU 1 lagerkort og Basalt I/O
- 2 maskine med 2 CPU 2 lagerkort og Basalt I/O

Kortpriser for prototyper.

Der laves 2 serier af prototyper for hvert af de 3 kort (CPU, lager basalt I/O).

CPU			
	Komponenter:	10 * 50.000	500.000
	Print:	2 * 60.000	120.000
Lager			
	Komponenter:	10 * 50.000	500.000
	Print:	2 * 60.000	120.000
Basalt I/O			
	Komponenter:	6 * 20.000	120.000
	Print:	2 * 60.000	120.000
Backplane		6 * 30.000	180.000
			1.660.000
Kabinetter:			+ 500.000
Gate array:			+ 1.000.000
Testudstyr:			+ 300.000
			= 3.500.000

Herudover skal der udvikles Net- og Disk-I/O-kort. Dette sker efter den første base. Udviklingsomkostningerne pr. kort er ca. 250.000 kr.

Der er ikke estimeret hardware-omkostninger til udvikling af Gogolplex.

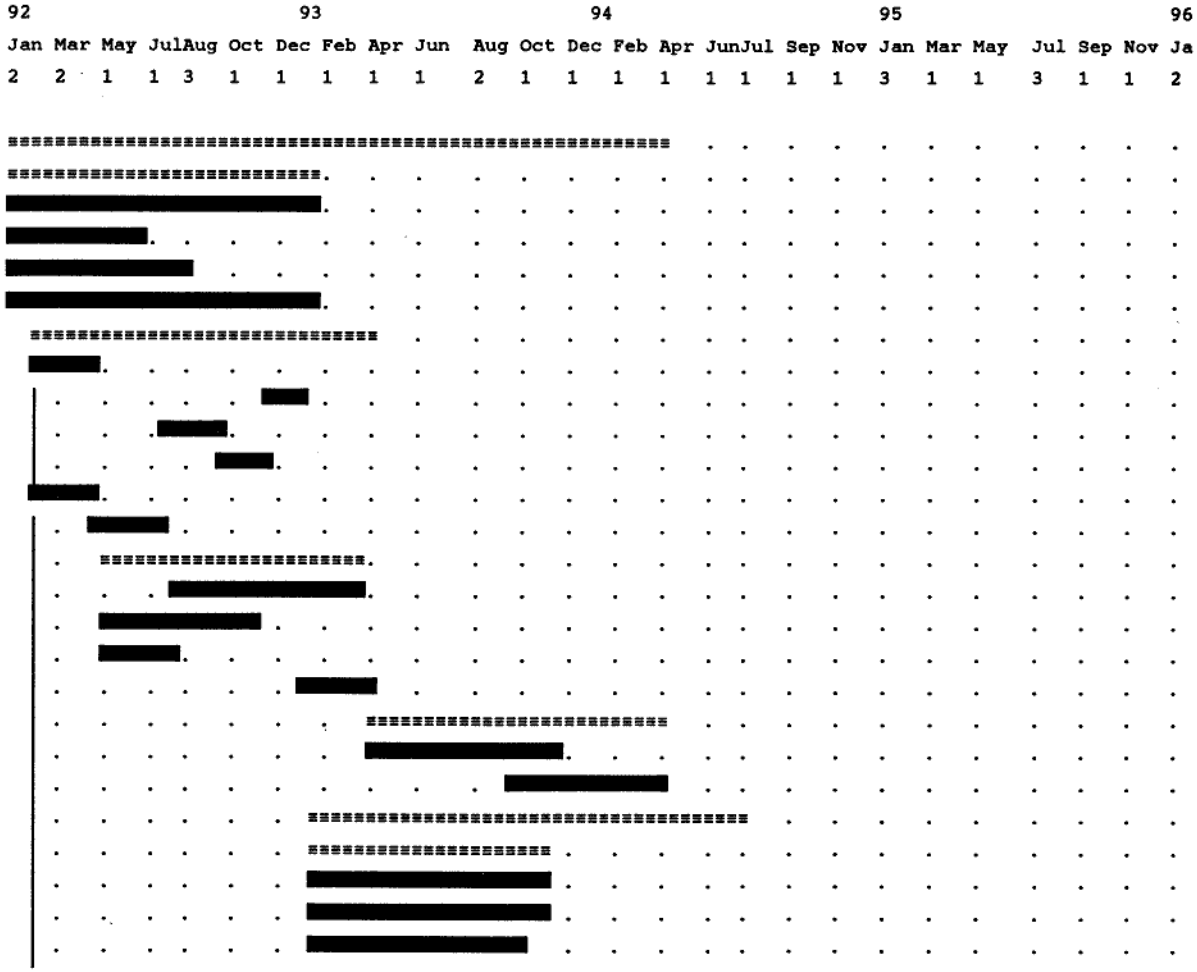
Indførelsen af et alternativt styresystem vil medføre en udgift på ca. 1.000.000 kr.

9. Projektplan

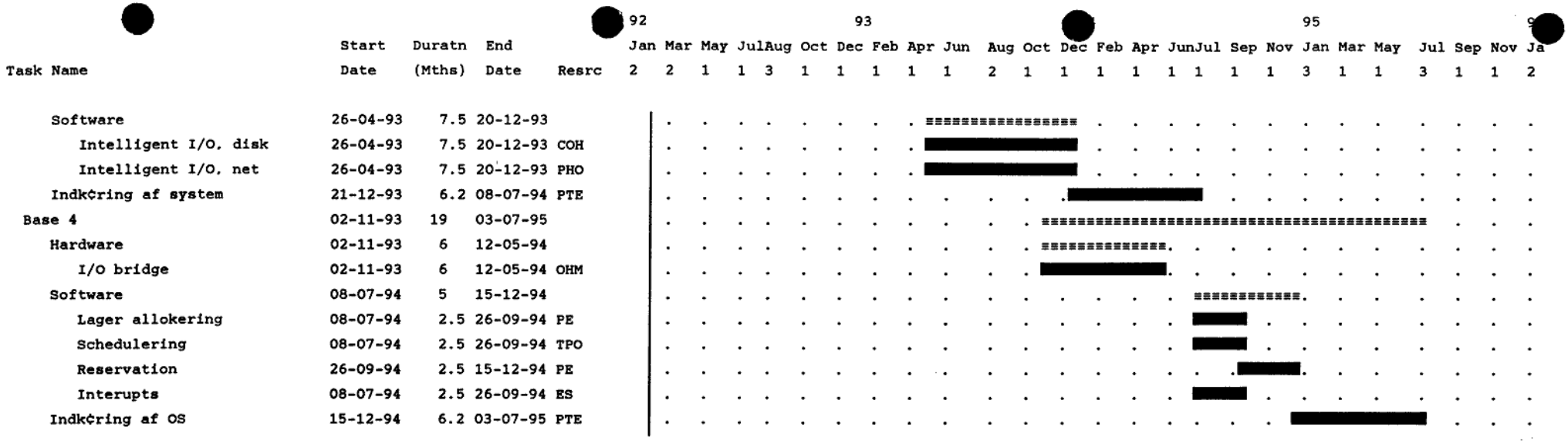
De følgende 4 sider giver en skitse til projektplan. Der er indlagt platforme på følgende baser:

Base II:	Ny bus	Leveres 30/3 94.
	R4000 CPU	
	Basal I/O	
	SMOS	
Base VI:	Ny bus	
	R4000 CPU	
	Basal I/O	
	Alternativ OS	
	I/O moduler	
	I/O lagerbro	
	Standard I/O	

Schedule Name : SPC/3 projekt, base 1,2,3 og 4
Responsible :
As-of Date : 03-02-92 Schedule File : BASE1-4'



Task Name	Start Date	Duratn (Mths)	End Date	Resrc
Base 1 og 2	02-01-92	25.6	30-03-94	
Hardware	02-01-92	12	20-01-93	
Businterface	02-01-92	12	20-01-93	KAN
CPU-modul	02-01-92	5	08-06-92	BS
Lager-modul	02-01-92	7	10-08-92	OHM
Basalt I/O	02-01-92	12	20-01-93	HW2
Software	03-02-92	13.1	26-03-93	
Flytning af Items	03-02-92	2.5	20-04-92	SW1
Globale data fra pl -> GM	24-11-92	1.3	06-01-93	PE
I/O via fællesbus (fit/free	09-07-92	2.4	21-09-92	PE
Net/Term I/O over streams	21-09-92	2	24-11-92	PE
Disk I/O over diskcache i G	03-02-92	2.5	20-04-92	PE
Ny interupt håndtering	21-04-92	2.5	08-07-92	PE
Basale I/O drivere	01-05-92	10	18-03-93	
Netdriver	21-07-92	7.5	18-03-93	ES
Diskdriver	01-05-92	6	06-11-92	TPO
Konsol driver	01-05-92	2.5	20-07-92	ES
R4000 TLB og lagersystem	06-01-93	2.5	26-03-93	PE
Indk¢ring og genering	26-03-93	11.5	30-03-94	
Indk¢ring af system	26-03-93	7.5	19-11-93	PTE
Platformsgenering	16-09-93	6	30-03-94	
Base 3	21-01-93	16.8	08-07-94	
Hardware	21-01-93	9	01-11-93	
I/O controller, disk	21-01-93	9	01-11-93	BS
I/O Controller,net	21-01-93	9	01-11-93	OHM
Standard I/O system	21-01-93	8	29-09-93	KAN



Detail Task ===== Summary Task ····· Baseline

(Progress) ==### (Progress) ▸▸▸ Conflict

(Slack) ###— (Slack) ..▓▓ Resource delay

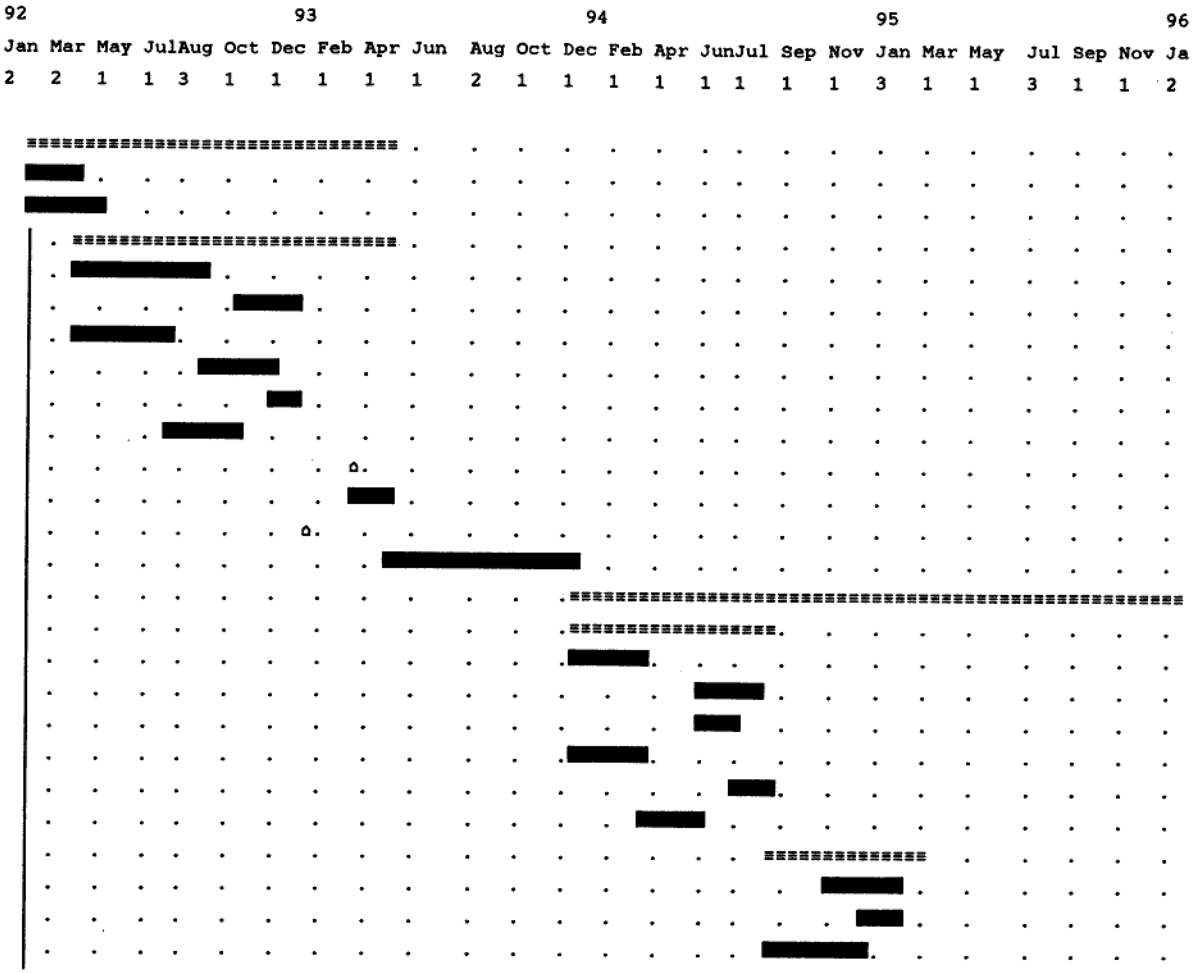
Progress shows Percent Achieved on Actual ◊ Milestone

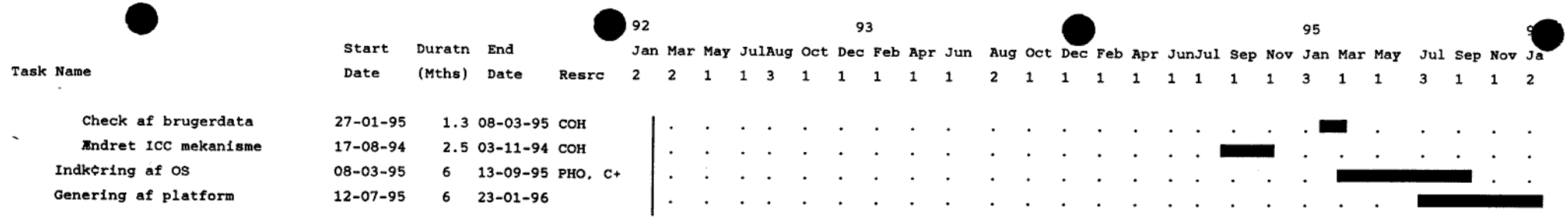
---------- Scale: 2 weeks per character ----------

TIME LINE Gantt Chart Report, Strip 1, Page 2

Schedule Name : SPC/3 projekt, base 5 og 6
Responsible :
As-of Date : 03-02-92 Schedule File : BASE5-6'

Task Name	Start Date	Duratn (Mths)	End Date	Resrc
Base 5, Standard SVR4/MP, UMA	03-02-92	14.1	27-04-93	
Indlæring	03-02-92	2	03-04-92	PHO, C+
Bootstrapping prom kode	03-02-92	3	05-05-92	
HW-afhængige dele	06-04-92	12.1	27-04-93	
Execption håndtering	06-04-92	5	09-09-92	COH
Interupthåndtering	20-10-92	2.5	11-01-93	PHO
Lagerhåndtering	06-04-92	3.7	31-07-92	PHO
Initialisering	10-09-92	2.5	30-11-92	COH
Big/little endian håndterin	01-12-92	1.3	11-01-93	COH
Håndtering af flydende kred	31-07-92	2.5	20-10-92	PHO
I/O drivere færdige (SMOS)	19-03-93	0	19-03-93	
I/O driver tilpasning (SMOS	19-03-93	1.3	27-04-93	PHO
Hardware tilgængelig	21-01-93	0	21-01-93	
Intelligente I/O controllere	26-04-93	7.5	20-12-93	
Base 6, NUMA	21-12-93	23.8	23-01-96	
Trin 1 ...	21-12-93	7.5	16-08-94	
Lokal allokering, stack ...	21-12-93	2.5	11-03-94	PHO
Rettelse af swap	31-05-94	2	01-08-94	PHO
Process-udvælgelse	31-05-94	1.3	08-07-94	COH
Disk I/O fra global cache	21-12-93	2.5	11-03-94	COH
Rettelse af Proc-system	08-07-94	1.3	16-08-94	COH
Test	14-03-94	2.5	27-05-94	PHO, C+
Trin 2	17-08-94	6.2	08-03-95	
Opdeling lokal/global	04-11-94	2.5	26-01-95	COH
"OS stack" til lokalt lager	15-12-94	1.3	26-01-95	PHO
Flytning af MMU-træ lokalt	17-08-94	3.7	15-12-94	PHO





Task Name	Start Date	Duratn (Mths)	End Date	Resrc
Check af brugerdata	27-01-95	1.3	08-03-95	COH
Ændret ICC mekanisme	17-08-94	2.5	03-11-94	COH
Indk¢ring af OS	08-03-95	6	13-09-95	PHO, C+
Genering af platform	12-07-95	6	23-01-96	

Detail Task ≡≡≡≡≡ Summary Task ••••• Baseline
(Progress) ≡≡≡≡≡ (Progress) ▸▸▸ Conflict
(Slack) ≡≡≡— (Slack) .. Resource delay
Progress shows Percent Achieved on Actual ◊ Milestone

---------- Scale: 2 weeks per character ----------

10. KONKLUSION

Det har været hensigten med denne rapport, at få skabt en fælles referenceramme, dels blandt nuværende og kommende projektdeltagere, dels i direktionen. Dette letter kommunikationen og beslutningsprocessen i det kommende projektforløb.

Det er herefter nødvendigt at få fastlagt en overordnet kravspecifikation, som entydigt beskriver den kommende Supermax's funktionalitet og ydelse. De valg, der er beskrevet i denne rapport, vil først være besluttet og fastlagt, når kravspecifikationen er udarbejdet og godkendt.

For at kunne skrive denne kravspecifikation er det nødvendigt at lukke en række af de løse ender, som stadig findes, og som fremgår af denne rapport. Vi foreslår derfor et forløb, hvor følgende aktiviteter finder sted parallelt de næste par måneder:

- analyse af hvilke tiltag der er nødvendigt for at leve op til Main-frame miljøer's krav til 'high availability'.
- definition af back-plane bus.
- detaljeret projektplanlægning.
- udarbejdelse af overordnet kravspecifikation.