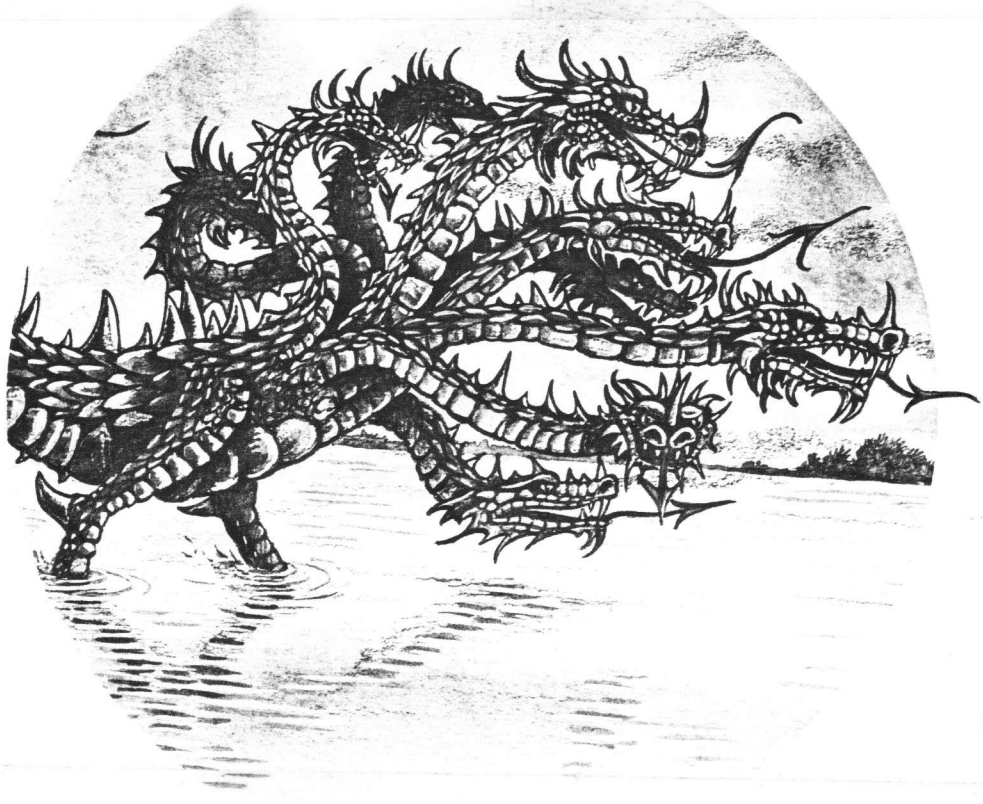
HYDRA

parallelle verwerking van langdurig rekenwerk in een netwerk met werkstations



Het oude Griekenland werd geteisterd door een kolossale draak, Hydra genaamd, die over negen koppen beschikte. Het middelste hoofd was onsterfelijk; werd een van de andere koppen afgeslagen, dan kwam er onmiddellijk een nieuwe voor in de plaats. ***(Uit: "Wonderwezens", Theo Schildkamp)***

Hydra, IC

doelstelling

- ***het definieren van een applicatiemodel waarbij***
  - ***applicaties herstartbaar zijn***
  - ***op verschillende machines parallel kunnen worden verwerkt***

- ***het bieden van een zo eenvoudig mogelijke overgang van een niet-parallelle applicatie naar een parallelle applicatie***

- ***het bieden van een besturingsmechanisme waarmee synchronisatie en voortgang van de parallelle applicatie volledig wordt geregeld.***

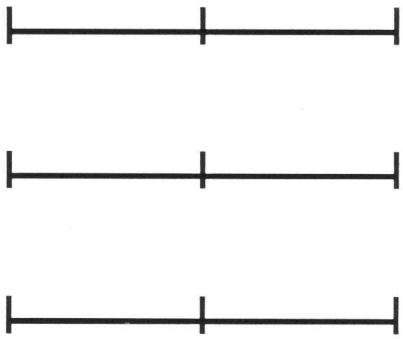
Applicatie model

langdurig rekenwerk moet herstartbaar zijn

bijvoorbeeld: totale aantal iteraties opdelen in stappen:

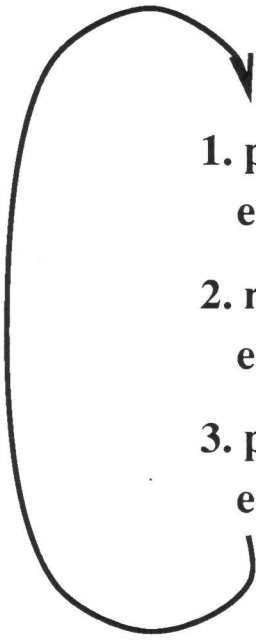
sommige applicaties kunnen opgesplitst worden in afzonderlijke delen, die min of meer onafhankelijk kunnen worden verwerkt

bijvoorbeeld: te verwerken invoer opsplitsen:



Applicatie model

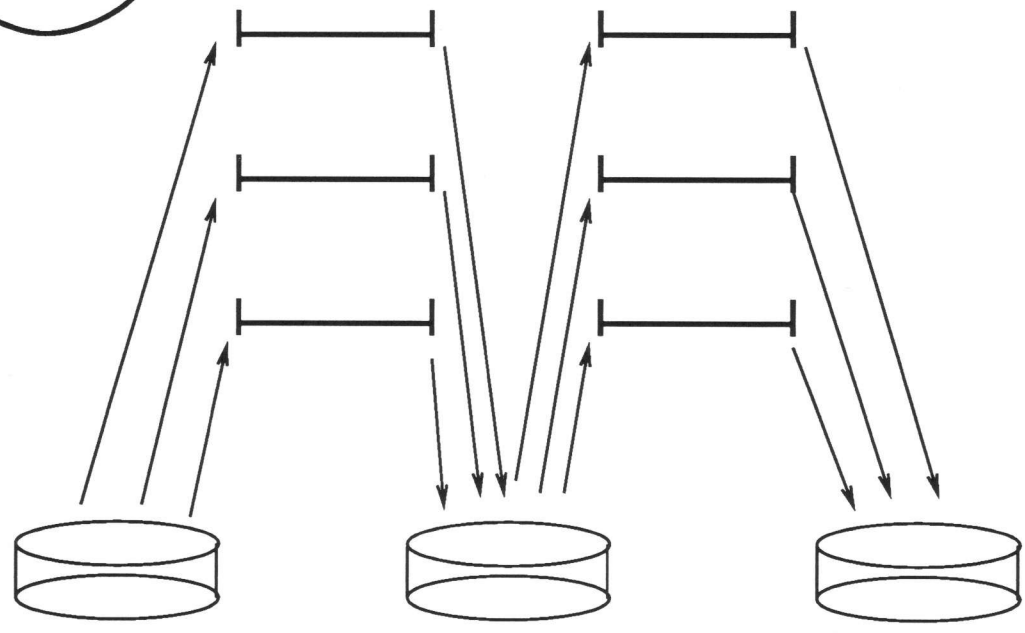
herstartbare parallelle applicaties kennen een cyclus:



1. parameters inlezen
eventueel gegevens inlezen

2. rekenen
eventueel lokale gegevens inlezen en wegschrijven

3. parameters voor volgende cyclus wegschrijven
eventueel globale gegevens wegschrijven



onderdelen van het pakket

MCP *Het 'Master Control Program' bestuurt de applicaties en zorgt voor error recovery en het verplaatsen van applicaties*

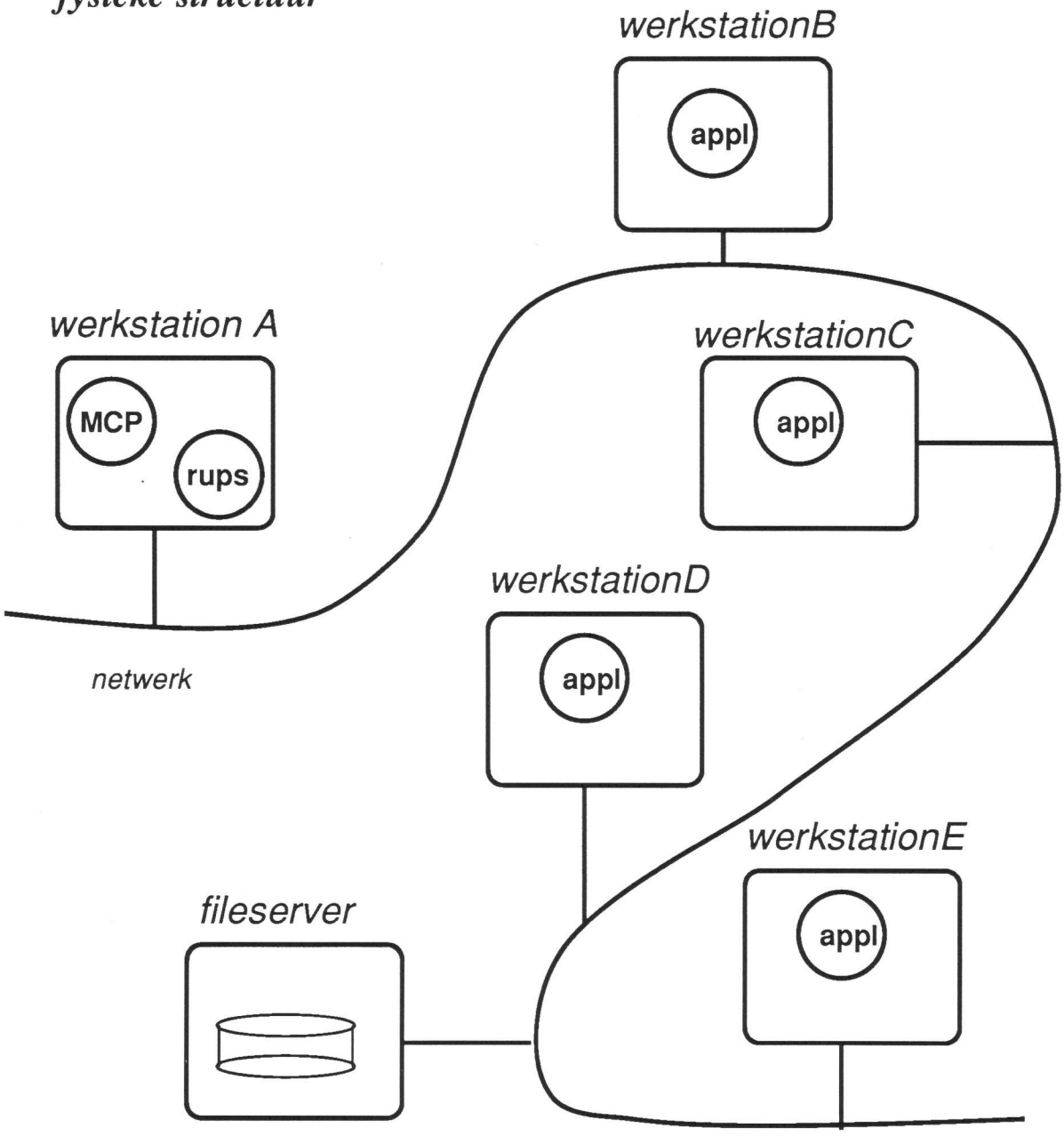
RUPS *De 'Remote Uptime Server' voorziet MCP periodiek van de gemeten "loadfactor" van alle hosts in het netwerk, die mogelijk gebruikt kunnen worden.*

ACP *Het 'Application Control Program' zorgt voor het opbouwen van een "TCP/IP socket" en het opstarten van de applicatie in een unieke directory.*

APPL *De gebruikers applicatie.*

Architectuur

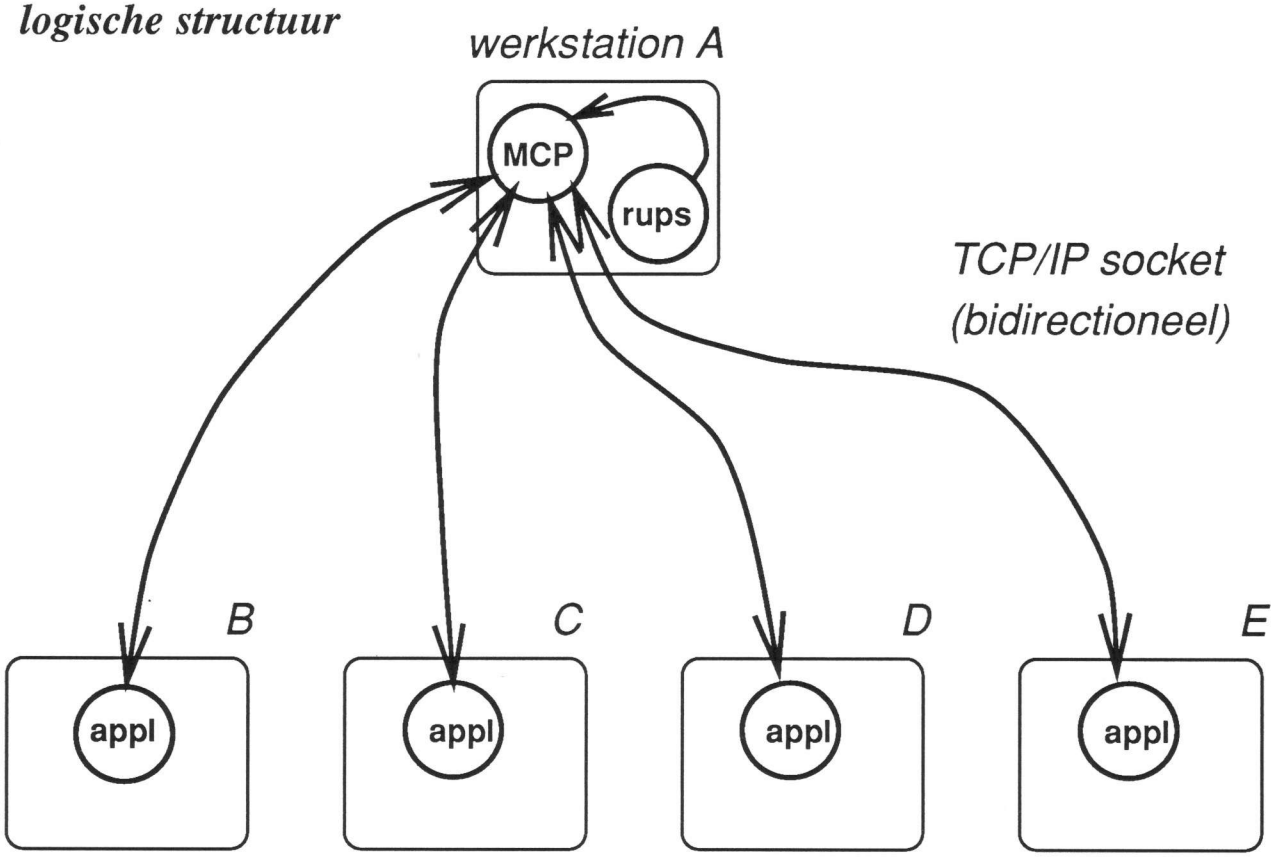
fysieke structuur



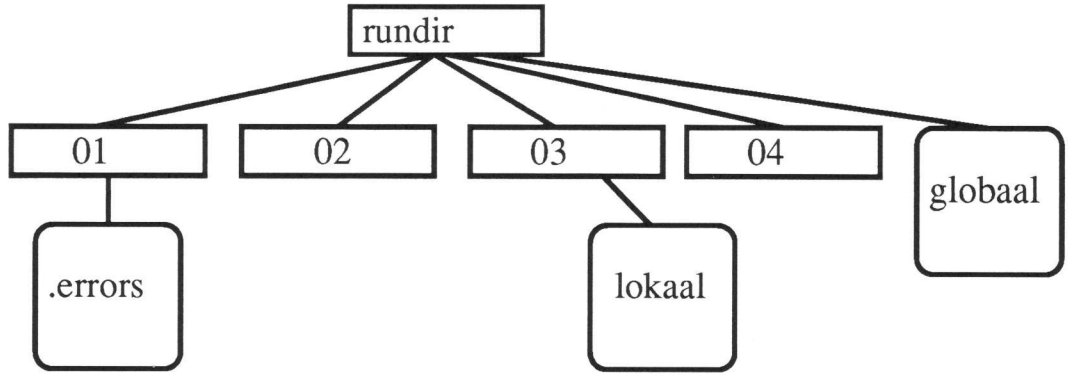
Hydra, IC

Architectuur

logische structuur



directory structuur:

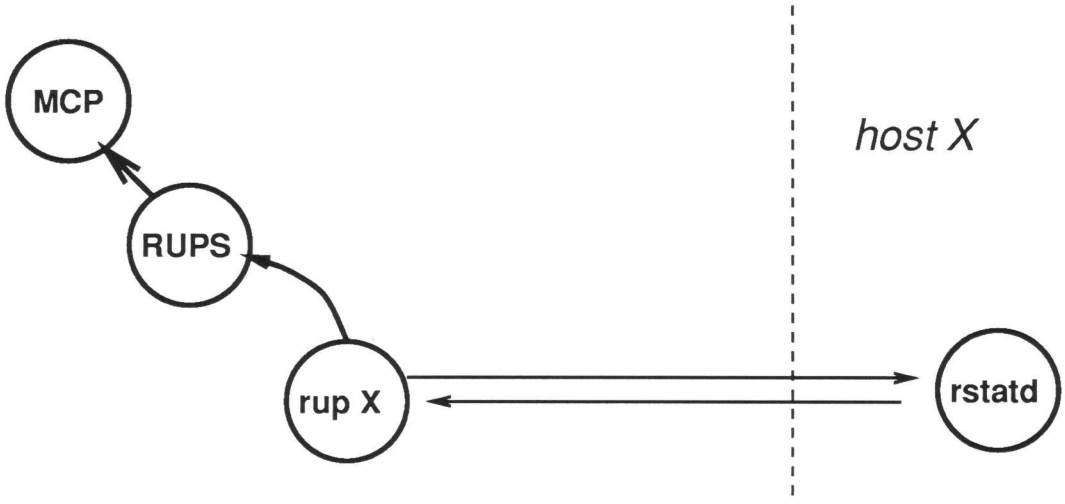


Hydra, IC

functies

- **kan n >= 1 applicaties besturen in een netwerk met werkstations**
- **verplaatst applicaties naar werkstations die lager belast zijn**
- **start applicaties opnieuw op een ander werkstation als het oorspronkelijke werkstation gecrashed is**
- **kan snellere typen werkstations onderscheiden van langzamere**
- **stopt (optioneel) de verwerking tijdens kantooruren**
- **stopt (optioneel) de verwerking als werkstations boven een instelbare belasting zijn gekomen en er geen beschikbaar zijn met een belasting onder die instelbare waarde**
- **kan in de 'schrijf' fase naar keuze synchroniseren: alle applicaties tegelijk ("simultaan") of een-voor-een**
- **kent een simpel protocol voor het aansturen van applicaties**

het meten van de belasting



RUPS: *Werkt de hostlijst in volgorde af. Draait voor iedere host het "rup" commando. Haalt uit de uitvoer van "rup" de actuele belasting. Als rup niets oplevert na een time-out wordt de host als "down" beschouwd.*
Nadat alle hosts benaderd zijn gaan alle "loadfactoren" in een pakket naar MCP. Vervolgens wacht RUPS een instelbare periode van maximaal 45 seconden.

MCP: *Verwerkt de ontvangen "loadfactoren". Bouwt een nieuwe lijst van vrije hosts op en berekent voor hosts waarop een applicatie loopt in de 'rekenfase' de gemiddelde load gedurende het rekenwerk.*
Beschouwt hosts waarvan al een tweede keer geen load-factor is ontvangen als "down" en herstart de applicatie op een andere host.

verplaatsing van een applicatie

- ***Applicaties worden verplaatst als hun elapsed-tijd meer dan 10 % afwijkt van het gemiddelde.***
- ***Applicaties worden verplaatst als de eerstvolgende vrije host verwacht wordt 10 % sneller te zijn.***
- ***(optioneel:) Applicaties worden verplaatst als de gemiddelde load tijdens de rekenfase een instelbare drempelwaarde overschrijdt.***

Optioneel kan de verwerking worden gestopt om twee mogelijke redenen:

TIMESUSPEND:

Het einde van de vorige cyclus valt binnen kantoortijd. RUPS wordt gestopt en 5 minuten voor het einde van de kantoortijd opnieuw gestart.

LOADSUSPEND:

Het aantal hosts waarvan de gemiddelde load tijdens de rekenfase groter is dan de drempelwaarde + 1.0 is groter dan het aantal vrije hosts met een lagere load dan de (instelbare) drempelwaarde.

loadfactor en ervaringscijfer

De "loadfactor" is een meetbare maat voor de belasting van een machine

load = 3.0

Gemiddeld 3 processen houden in de afgelopen periode de CPU bezet.

Het "ervaringscijfer" is een berekende schatting van de tijd die een applicatie er over doet als de betreffende host leeg zou zijn.

$$\text{ervaringscijfer} = \frac{\text{elapsed tijd van de rekenfase}}{\text{gemiddelde loadfactor in rekenfase}}$$

load = 3.0 **elapsed tijd = 15 minuten**

ervaringscijfer = 5 minuten

Sortering in vrije lijst op:

(actuele load + 1.0) * ervaringscijfer

Hydra, IC

loadfactor en ervaringscijfer

aanpassingen voor hosts met meerdere processoren

MAXLOAD= 0.3 (onbelast)

minimale aandeel in de belasting: $\frac{1}{1 + 0.3} = 77\ \%$

De 'wn3' geeft een loadfactor voor alle processoren (4 stuks) samen.

Als we een maximale load van 1.3 willen hanteren, dan willen we minimaal 77% per processor.

Het aantal processen, dat we toelaten op de 'wn3' wordt iteratief bepaald:

#cpu = aantal cpu's = 4

Load = loadfactor = bijv. 3.0

p = aantal toe te laten processen

MAXLOAD = instelbare drempel = 0.3

voor p van 1 tot en met #cpu

$$\text{Als} \quad \#cpu * \frac{1}{Load + p} > \frac{1}{1 + MAXLOAD}$$

dan mag p toegelaten worden

Bij een load van 3.0 worden twee processen toegelaten.

Bij een load van 2.0 zijn het er drie.

Hydra, IC

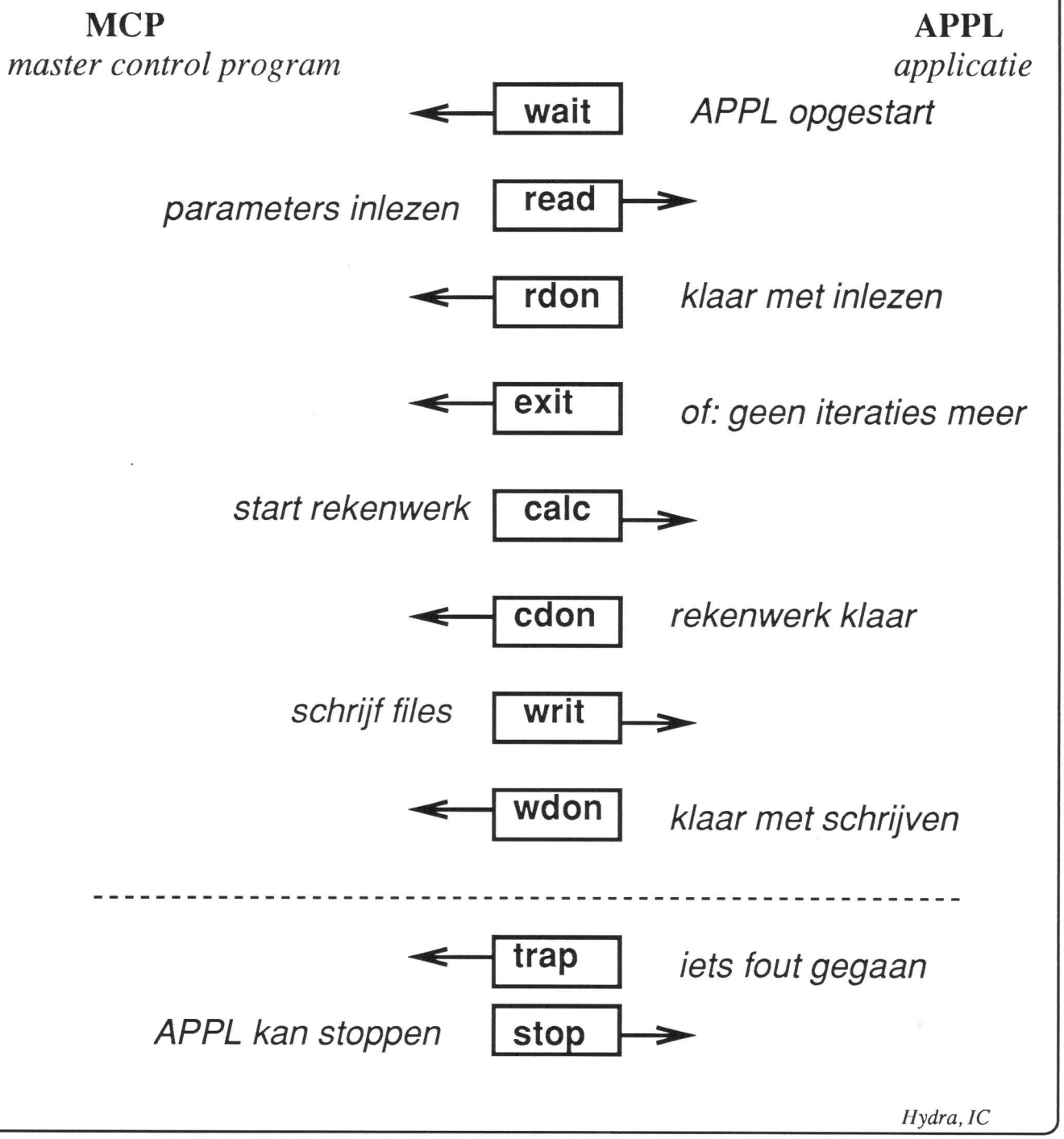
status display

```
MCP v5 running 8 '../testapl'  cycle #4   MS_CALC   [ 32:28 ]

# host       load                        status     remarks
-- ---------- --------------------------- ---------  -------
01 groen      0.18 <                      AS_CALC
02 geel       0.83 >>>>                   AS_CALC
03 indigo     0.59 <<<                    AS_CDON    01:10 elapsed
04 oker       0.91 |||||                  AS_CDON    01:11 elapsed
05 schwarz    1.08 >>>>>>                 AS_CALC
06 bordeaux   0.73 <<<<                   AS_CDON    01:12 elapsed
07 cyaan      0.59 <<<                    AS_CALC
08 oranje     0.92 >>>>>                  AS_CDON    01:19 elapsed


azuur   0.23/72      grijs   0.08/86      blanc   0.09/88
noir    0.05/93      wit     0.16/88      gris    0.13/91
gray    0.28/87      black   0.36/91
```

logfile

```
26/5 4:08:55 start cycle 36

26/5 4:13:11 appl #16 crashed on host 'wn2'

26/5 4:13:41 Caught timeout invoking #16 on 'gris'

26/5 4:14:07 Restarted #16 on weiss

26/5 4:14:32 Host 'wn2' considered down

26/5 4:30:09 end cycle 36, 15:59 elapsed

 ---------------------------------

26/5 6:21:06 start cycle 44

26/5 6:25:22 Host 'wn2' alive

26/5 6:35:16 end cycle 44, 14:10 elapsed

26/5 6:35:55 Moved #11 from gris to wn2

26/5 6:35:59 start cycle 45
```

Protocol


MCP
master control program
APPL
applicatie
wait
APPL opgestart
parameters inlezen
read
rdon
klaar met inlezen
exit
of: geen iteraties meer
start rekenwerk
calc
cdon
rekenwerk klaar
schrijf files
writ
wdon
klaar met schrijven
trap
iets fout gegaan
APPL kan stoppen
stop

protocol ondersteuning

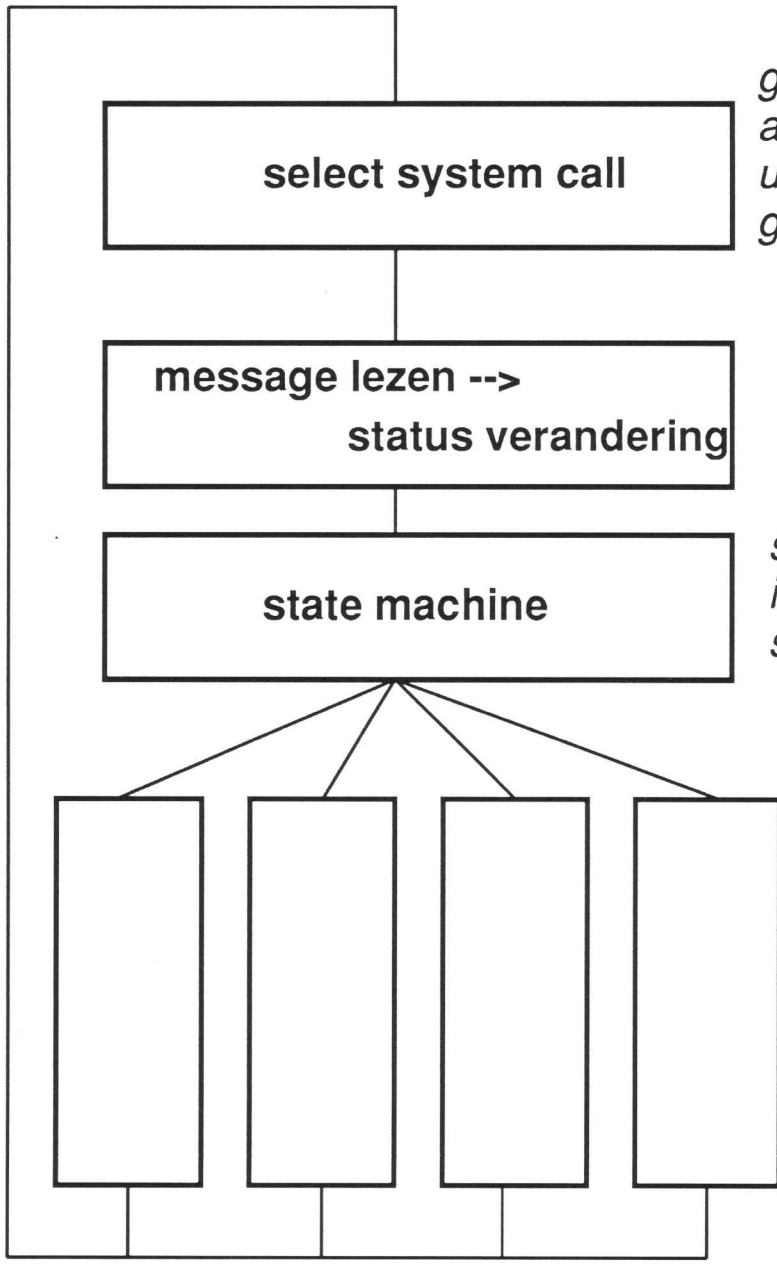
(include file en routines voor Fortran en C)

```
#include "mcpprotocol.h"

sendm(M_WAIT);

mes = receivem();
switch( mes ) {
  case M_CALC:
          . . .
  case M_WRIT:
          . . .
```

```
#include "mcpfprot.h"
      call fsendm( M_WAIT )
      call frecvm( mes )
      if( mes .eq. M_CALC) then
                . . .
      else if( mes .eq. M_WRIT) then
                . . .
```

Implementatie

basiscyclus van MCP



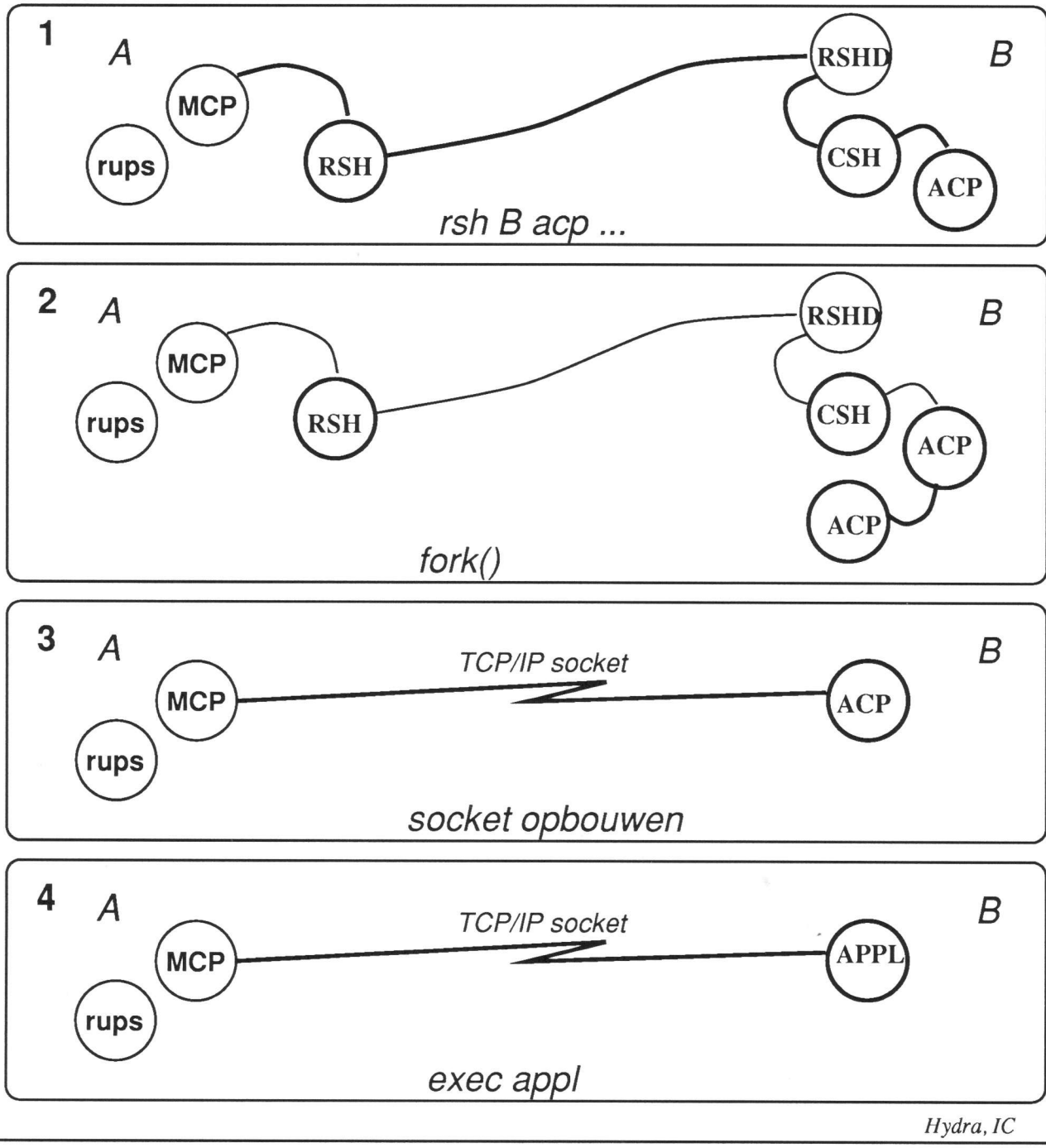
geeft bitmasker, waarin aangegeven welke files uitvoer hebben, die gelezen kan worden

status verandering is ingang voor statustabel

modules handelen specifieke status overgangen af

Architectuur

opstart scenario

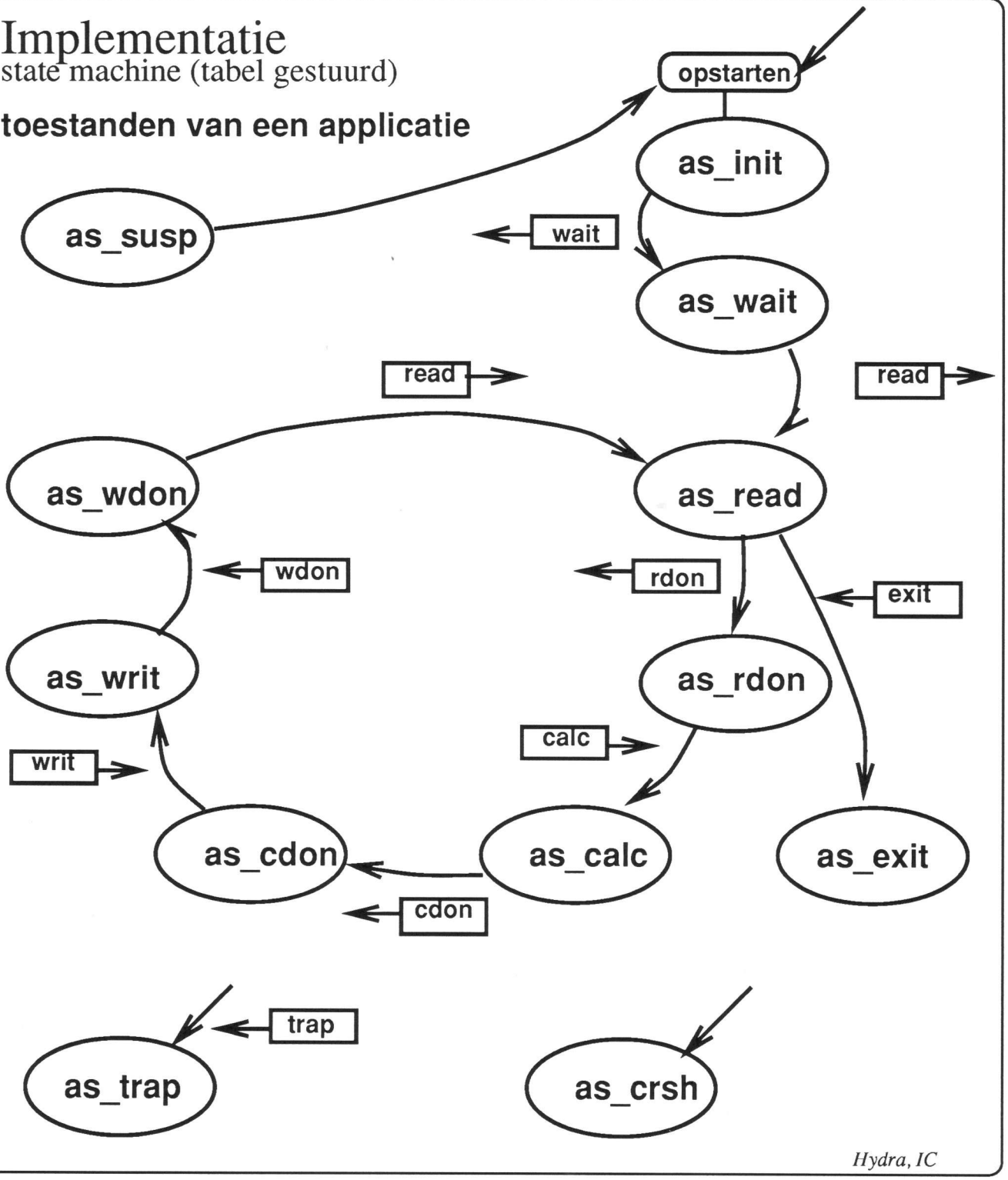


Hydra, IC

Implementatie

state machine (tabel gestuurd)

toestanden van een applicatie

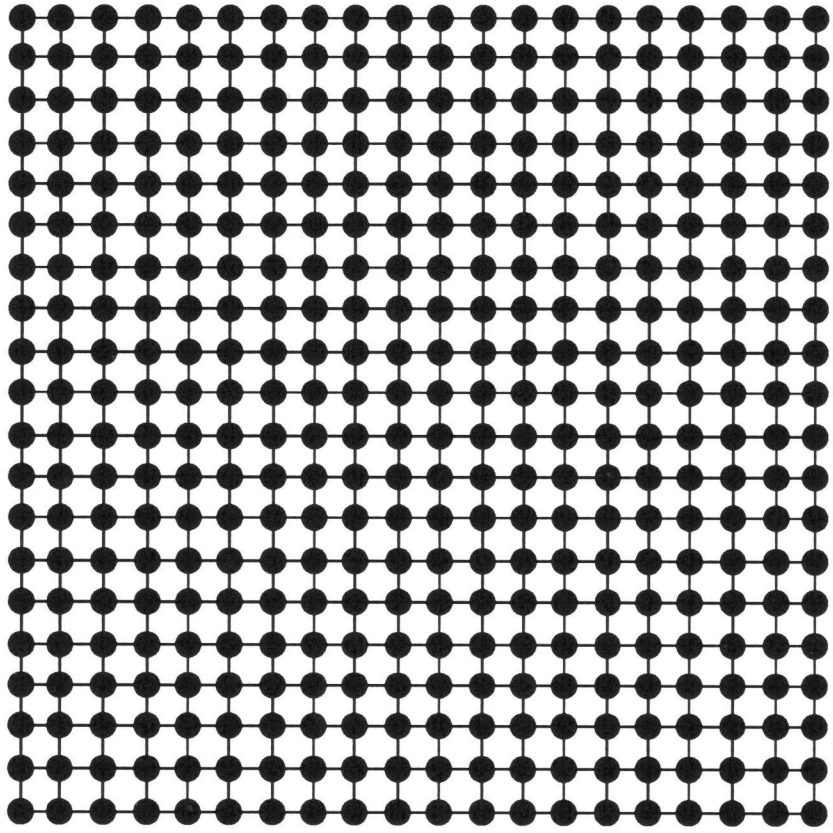


Hydra, IC

een toepassing

parallel Kohonen netwerk (Melssen / Smits)

Neuraal netwerk van 20 bij 20 neuronen. Te verwerken leerset bestond uit 3284 infrarood spectra.



Totaal 5.000 'cycles' in zes weken op twaalf machines.

Rekentijd op een Sun 4/40 IPC: > 450 dagen.