Bilag til OLA med RMnet

Snitfladespecifikationer

Dokumentets placering	X:_Begraenset\Netvaerk\RmNet\OLA\Snitfladespecifikationer.pdf
Version	alpha-3
Seneste revision	2010-01-20 af peter.rathlev@stab.rm.dk

Indholdsfortegnelse

1 Introduktion

Dette dokument forsøger så præcist som muligt at beskrive de snitflader der er mellem på den ene side endepunkter ("hosts") og regionsnetværket og på den anden side mellem autonome netværk (f.eks. et lokalnet) og regionsnetværket.

Beskrivelserne er tekniske og derfor primært henvendt til teknikere. Princippet for beskrivelserne er at samle hvad der allerede nu er etableret og virker. Dokumentet forsøger derfor ikke overlagt at introducere nye standarder eller procedurer for hvordan netværk i Region Midtjylland benyttes.

Dokumentet indeholder ingen følsomme oplysninger, og er i høj grad også henvendt til nuværende eller potentielle leverandører af IT-systemer til Region Midtjylland.

Bemærk: Dokumentet er for øjeblikket i "alpha"-revision, og skal altså mere ses som vejledende end dikterende. Vi modtager meget gerne input til forbedringer eller præciseringer.

2 Generelle anbefalinger

Udover de i senere kapitler nævnte konkrete specifikationer bør den driftsansvarlige part over for RMnet tage følgende til efterretning.

2.1 Redundans

Som udgangspunkt leveres alle forbindelser, både til endepunkter og til netværksudstyr, som redundante forbindelser.

- Det anbefales at **teste eventuel failover** (f.eks. ved at trække kabler ud af porte) mellem eventuelle flere redundante forbindelser.
- Hvis det er muligt at angive hvordan et *endepunkt* overvåger om et link blandt flere redundante kan bruges anbefaler vi "MII monitoring"-metoden, der løbende kontrollerer om der er "link up". For mange systemer er det muligt at benytte ARP- eller ICMP-probes til at teste, og omend dette i visse scenarier giver et mere pålideligt resultat resulterer det i en væsentlig belastning af netværket.
- *Redundans er den enkelte hosts ansvar*. Hvis der leveres flere forbindelser til en given host (typisk 2 for simpel redundans) er det hostens ansvar at opdage et inaktivt link og aktivere et andet i stedet. Det kan i den forbindelse antages at et aktivt link (dvs. hvor der kan forhandles "link up") vil være i stand til at bære trafik; dette er sikret ved at access-switchene benytter "link poisoning" der automatisk lukker downlinks hvis der ikke er nogen brugbare uplinks.
- Fra den enkelte access-switch og opefter er netværket som udgangspunkt opbygget med simpel redundans som minimum, dvs. uden noget "single point of failure". I forbindelse med uplanlagte hændelser vil netværket som udgangspunkt kunne "konvergere" (dvs. arbejde sig uden om den opståede fejl) i løbet af 3 minutter.

2.2 Performance

Det er for det meste nødvendigt at tage nogle ting i betragtning når man ønsker at optimere performance for systemer der benytter netværket. Dette gælder specielt for centralt placerede systemer i Region Midtjylland, der ofte tilgås fra over 100 km væk.

- På grund af de relativt store afstande i Region Midtjylland anbefales det at teste performance grundigt inden idriftssættelse, specielt for systemer hvor de enkelte komponenter står langt fra hinanden. (Eksempelvis på to forskellige driftscentre eller på to forskellige sikkerhedsniveauer.) Som udgangspunkt skal en **latens på 20 ms** forventes inden for Region Midtjyllands eget net. Hvis der er behov for garantier for lavere latens end dette mod et givent punkt skal det aftales på forhånd.
- Vi stiller som udgangspunkt **ikke nogen båndbreddegarantier**. Hvis det ønskes skal RMnet gerne lave en båndbreddetest mellem to endepunkter i nettet. Det anbefales ***ikke*** at forsøge sig med f.eks. filoverførsler mellem servere, da dette meget sjældent giver et retvisende billede af den båndbredde der er tilgængelig i netværket.
- Da mange operativsystemer benytter en række meget konservative default-indstillinger for især TCP-stakkens konfiguration anbefales det kraftigt at foretage relevante justeringer ifølge "best

practices[1]" for store højhastighedsnetværk, såkaldte "**long fat networks**". Bemærk at TCP-protokollen netop er designet til at være hensynsfuld over for resten af netværket, og at maksimal performance derfor naturligt begrænses når TCP benyttes. Af den grund vil en almindelig hastighedstest over en ikke-justeret TCP-forbindelse næsten altid give et dårligere resultat end hvad netværket vil kunne levere.
Det anbefales specifikt at sørge for at **TCP Window Scaling** er slået til, og at TCP receive buffers er tilstrækkeligt store. Et endepunkt der benytter auto-tuning er optimalt; alternativt bør **TCP receive buffers være mindst 1 MB**, hævet fra udgangspunktet på 64 kB.

- Maksimum **IP MTU er 1500 bytes**. For trafik der skal krydse de centrale firewalls må en fragmenteret IP-pakke efter "reassembly" derudover maksimum være 8782 bytes. Er pakken større end dette vil den blive droppet[2]. Hvis der er behov for at transportere data der ikke følger denne profil skal det aftales med RMnet på forhånd.
- Som udgangspunkt behandles al trafik ens[3]. Hvis der er behov for specielle trafik-prioriteringer kan dette evt. aftales med RMnet.

2.3 Andet

- *Stabilitet*: Ofte vil trafik gennem regionsnetværket krydse en eller flere stateful firewalls. For at undgå at disse firewalls lukker inaktive forbindelser bør alle forbindelser benytte **TCP Keep-Alives**, hvis de på den ene side ønskes opretholdt gennem længere tid (mere end 2 timer) og på den anden side ikke nødvendigvis er aktive hele tiden. Regionsnetværket overholder RFC 5382[4], "NAT Behavioral Requirements for TCP".
- *Fejlfinding*: Vi vil opfordre til at der altid udarbejdes en **kommunikationsoversigt** for de enkelte endesystemer. Denne skal så præcist og udførligt som muligt beskrive hvordan der kommunikeres med andre systemer. Beskrivelsen forventes at kunne bruges til f.eks. at lave detaljeret trafikfiltrering hvis det skulle skønnes nødvendigt.
- *Features*: Regionsnetværket router for øjeblikket **udelukkende unicast IPv4-trafik**. Specifikt er IPv4 multicast og IPv6 ikke understøttet.
- *Baseline*: Det anbefales at foretage baselining inden systemer idriftsættes. En egentlig "**Operational Acceptance Test**" er at foretrække, og kan bruges til at vurdere senere mistanker om dårlig performance.

1 Se f.eks. "TCP Extensions for High Performance", http://tools.ietf.org/html/rfc1323
2 http://www.cisco.com/en/US/docs/security/fwsm/fwsm31/configuration/guide/specs_f.html#wp1055440
3 Dog med undtagelse af voice-trafik der behandles som "Expedited Forwarding"-trafik.
4 http://tools.ietf.org/html/rfc5382

3 Fysiske specifikationer

Alle enheder der sluttes til regionsnetværket forventes at leve op til de i dette afsnit angivne fysiske specifikationer.

3.1 Fysiske specifikationer for endepunkter

I dokumentet er udtrykket "host" brugt om endepunkterne; dette kan være både servere og andre enheder der benytter netværket. Fælles for disse endepunkter er at de i sig selv ikke leverer forbindelse til andre enheder[5].

- Alle forbindelser leveres via RJ45 8P8C (kobber).
- Den præcise fysiske snitflade vil være enden af et UTP-kabel.
- Det benyttes som minimum UTP Cat5e-kabling; det skal forventes at alle 8 ledere er i brug, f.eks. til 1000BaseT.
- Den fysiske forbindelse følger IEEE 802.3ab-standarden[6]. Gigabit-forbindelser ***skal*** benytte auto-negotiation.
- Hvis en host har behov for et 100Mb- eller 10Mb-link skal dette aftales på forhånd. I dette tilfælde anbefales det ***ikke*** at benytte auto-negotiation.

3.2 Fysiske specifikationer for netværksudstyr

Netværksudstyr betyder i denne kontekst en hvilken som helst enhed der selv leverer forbindelse til andre enheder.

- Alle links fungerer som Gigabit-links, dvs. hastigheder højere eller lavere end dette som udgangspunkt ikke leveres.
- Forbindelsen kan enten leveres som 1000BaseT/TX (kobber) eller som lyslederforbindelse over enten multimode-fiber (850 nm eller 1300 nm transceivere, Cisco "SX") eller singlemode-fiber (1310 nm eller 1550 nm transceivere, Cisco "LX/LH").
- Den præcise fysiske snitflade vil være enden af et UTP-kabel eller enden af et MM- eller SM-kabel med enten SC- eller LC-connector; det forventes at endesystemet selv indeholder en relevant transceiver.

5 Virtualisering, f.eks. vha. VMware eller XEN, ligger i en gråzone.
6 IEEE 802.3-2008 section 3, http://standards.ieee.org/getieee802/802.3.html

4 Logiske specifikationer

Alle enheder der sluttes til regionsnetværket forventes at leve op til de i dette afsnit angivne logiske specifikationer. Hvis der er behov for afvigelser fra de her angivne parametre skal der træffes forudgående aftale med RMnet.

4.1 Logiske specifikationer for endepunkter

For endepunkter, dvs. enheder der ikke leverer forbindelse til andre endepunkter, leveres forbindelserne med følgende logiske specifikationer:

- Switchportene accepterer som udgangspunkt udelukkende "untagged" trafik, dvs. almindelig Ethernet II[7] framing. Specifikt accepteres som udgangspunkt ikke IEEE 802.1Q-trafik. Hvis der er behov for dette skal der træffes særskilt aftale.
- *Hvis der er behov for at fremføre mere end et enkelt VLAN* til en host vil dette ske ved hjælp af IEEE 802.1Q trunks. Dette vil altid skulle aftales på forhånd. Et sådant trunk vil altid være manuelt konfigureret, dvs. der foretages ingen dynamisk forhandling.
- Kun Ethertypes 0x0800 (IPv4) og 0x0806 (ARP) kan forventes at blive behandlet korrekt.
- Som udgangspunkt accepteres kun en enkelt MAC-adresse per port. RMnet vil meget gerne have besked hvis en host forventes at sende fra flere MAC-adresser[8].
- Alle porte er konfigurererede til at afvise STP BPDU'er, og et access-link vil automatisk blive lukket hvis switchporten modtager BPDU'er. Dette gælder også for 802.1Q trunks.
- Af de to foregående punkter følger at et endepunkt aldrig må bridge mellem flere fysiske porte. Dette skal specielt holdes for øje når endepunktet selv er i stand til at levere switch-funktionalitet.
- Det er kun tilladt at bruge IP-adresser der er blevet tildelt hosten eksplicit, jvf. gældende procedurer[9].
- Hvis der er behov for mere båndbredde end hvad en Gigabit-forbindelse kan levere er det efter forudgående aftale muligt at benytte LACP (IEEE 802.3ad) link-aggregering. Bemærk at dette leveres uafhængigt af redundans; hvis man ønsker 2x1 Gb redundant forbindelse skal man bruge fire netkort i hosten.

4.2 Logiske specifikationer for netværksudstyr

Udgangspunktet er det samme som i kapitel 4.1 men med følgende justeringer:

- Links leveres som udgangspunkt som IEEE 802.1Q trunks. Som udgangspunkt accepterer vi ikke "untagged frames" (f.eks. "native vlan" i Cisco-terminologi) da disse uden 802.1Q-headeren ikke bærer CoS-modellens 802.1p bits til Quality of Service.
- Der benyttes ingen dynamiske forhandlinger på trunks (f.eks. Ciscos DTP), dvs. alle trunks

7 http://en.wikipedia.org/wiki/Ethernet_II_framing
8 Dette kan f.eks. være en VMware-server med indbygget "vswitch".
9 Tildelingen af IP-adresser i driftscentre varetages pt. af Server Management.

konfigureres eksplicit som trunks.

- Fra regionsnetværkets side er STP aktivt på disse links. Alle regionsnetenheder benytter Ciscos "Rapid PVST+"-model. For lokalt Cisco-udstyr anbefales dette ("`spanning-tree mode rapid-pvst`"). Alternativt vil regionsnetudstyret falde tilbage på IEEE 802.1w ("Rapid Spanning Tree") og derefter "klassisk" 802.1D spanning-tree.
- Som udgangspunkt forventes det at trafik sendt til regionsnetværket allerede er QoS-markeret. Efter aftale kan vi enten acceptere CoS- eller DSCP-markering. Hvis intet andet angives vil vi reklassificere al trafik til "Best Effort"-klassen (CoS 0/DSCP 0).
- QoS-snitfladen fra regionsnetværket, der er et MPLS VPN-netværk, er "Uniform Mode"; vi bærer altså al IP-trafik med samme QoS-markering som vi modtager trafikken med.
- Hvis der er en L3-snitflade mellem regionsnetværket og det lokale netværket er der mulighed for at benytte en dynamisk routnings-protokol mellem mellem de lokale enheder og regionsnetsudstyret. Vi understøtter OSPF (v2), RIP og BGP.
- Hvis ikke der benyttes en dynamisk routningsprotokol leveres redundansen mod regionsnettet vha. Cisco HSRP. Regionsnetværket benytter HSRP-grupperne 1-63, og andre HSRP-enheder på samme VLAN skal altså enten benytte gruppe 0 (den implicitte gruppe) eller et gruppenummer over 63. Vi benytter statisk routning mod lokalnettet og anbefaler at der også på den side etableres HSRP eller tilsvarende (VRRP/GLBP).
- Hvis der er behov for mere båndbredde end hvad et enkelt Gigabit-link kan levere er udgangspunktet at båndbredden øges vha. flere links i ECMP[10]-konfiguration. Her er det en fordel hvis der benyttes en dynamisk routningsprotokol mellem netværksudstyret og regionsnetnværket.

10 ECMP = Equal Cost Multi Path

5 Transport og filtrering

Ovenstående to kapitler dækker snitfladerne i kanten af regionsnetværket. Efter trafik er afleveret på kanten transporteres det gennem regionens core-netværk, og via en hierakisk struktur til en central firewall. Hvis trafik skal til eller fra Internettet vil det i øvrigt skulle passere en dedikeret firewall til dette.

Dette kapitel beskriver specifikationerne for disse.

5.1 Core-transport

Trafik der transporteres gennem core-netværket bliver som udgangspunkt underlagt RED[11] for at beskytte mod svær congestion og deraf følgende tail-drops og "savtakket" TCP-performance.

5.2 Centrale firewalls

Udgangspunktet for filtrering i regionsnetværket er "least privilege principle". Det betyder at som udgangspunkt er der lukket for al trafik, og at man eksplicit skal angive hvad der skal være adgang til og fra. Men kommunikationsplanen nævnt i kapitel 2.3 lettes dette betydeligt.

Der er en række ting der dog altid åbnes for, og som muliggør basal funktionalitet. Dette er:

- DNS-opslag mod de centrale DNS-resolvere (10.85.1.11, 10.85.1.12 og 10.83.1.11).
- NTP-synkronisering mod de centrale NTP-servere (aar-ntp.net.rm.dk og hor-ntp.net.rm.dk)
- ICMP Echo og Echo Reply (dvs. "ping")
- ICMP Time-Exceeded og Unreachables (f.eks. til traceroutes).

Bemærk at al trafik gennem firewallen logges til senere driftsanalyse.

5.2.1 Protocol inspection

Det kan ikke altid fra starten af specificeres hvad der skal åbnes for af kommunikation mellem to endepunkter. Visse protokoller forhandler først de specifikke detaljer for forbindelser over en kontrol-kanal når de skal bruges. Dette gælder bl.a. FTP, Oracle SQL*NET og visse RPC-baserede services, f.eks. Microsoft RPC eller standard Unix portmapper.

For at kunne åbne for de relevante forbindelse "on the fly" skal firewallen kigge detaljeret på kontrol-kanalerne. Dette løses i Cisco-terminologi via såkaldte "inspect" kommandoer. Udgangspunktet for regionsnetværket er ***at der ikke foretages inspection*** på denne måde af drifts- og informationssikkerhedsmæssige grunde. Hvis der er brug for at lave inspection af specifik trafik kan dette evt. aftales med RMnet.

5.3 Internet-firewall

Trafik mod Internettet er, som for den ovennævnte centrale firewall, som udgangspunkt lukket. Der er lavet en række standardåbninger som kan oplyses ved henvendelse.

11 Random Early Detection, dvs. at pakker smides væk med stigende sandsyndlighed efterhånden som buffers i netværket fyldes op.

Internetfirewallen foretager NAT[12] af forbindelser mod Internettet. Vi logger af driftsmæssige årsager alle disse oversættelser.

12 Network Address Translation