Information kring GleSYS driftavbrott 2012-07-07

Nedan följer information om vad som orsakade det driftavbrott som uppstod i Falkenberg och Stockholm Lördagen den 7:e Juli.

Bakgrund
Vårt nätverk sträcker sig numera över flera olika "siter". Varje site har en router som tar hand om trafiken som kommer till och från siten. I Falkenberg har vi två siter, i Stockholm har vi 3 st.
Alla dessa "siter" är sammankopplade så att de kan utbyta trafik.

Företag som GleSYS och dess kunder är ofta drabbade av sk. DDOS attacker. Det finns många sätt att bekämpa DDOS attacker och vi jobbar på flera fronter, men det finns ett sätt som är överlägset alla de andra. Det är nämligen så att "mest bandbredd vinner". Vi har därför satsat stort på att utöka vårt nät med så mycket bandbredd som möjligt. Därför har vi dessa "siter" där vi tar in internettrafik från olika internetleverantörer.

Vi drabbas dagligen emot DDOS attacker, ibland flera på samma gång, väldigt få av dessa drabbar våra kunder numera, just på grund av ovanstående lösningar.

Händelseförlopp, avbrott

13:49 - Vårt övervakningssystem varnar att en router i Falkenberg är överbelastad. Övriga tjänster börjar kort därefter ge larm om paketförluster i nätverket.

13:51 - Vi tappar helt kontaken med vår router över ordinarie förbindelse. *(Tekniker undersöker via sk "console" terminal på distans genom vårt interna backupnät.)*

14:05 - Tekniker utfärdar storlarm, och skickar instruktion till annan tekniker att fysisk närvaro på plats troligtvis kommer behövas.

14:25 - Tekniker på plats i serverhallen i Falkenberg

14:45 - En router i Falkenberg beter sig underligt och verkar ge minnesfel. Vi markerar den som trasig och stänger av. Routern är redundant, men viss redundans fungerar inte korrekt. Dessa fel rättas.
Trafik börjar fungera igen, men fortfarande mycket störningar.

14:55 Vi konstaterar att Stockholm inte blir hjälpt av rättningen av felet i Falkenberg. Tekniker kallas ut.

15:00 Vi beslutar att kalla in så mkt personal som möjligt. Svårt, eftersom det är semestertider och folk har en bit att åka. Bland annat så ligger interna system nere, vilket försvårar uppdatering av driftinfo på glesys.se som felaktigt visar att vi inte har några problem.

15:30 Trafik flyter bättre och bättre i Falkenberg. Stockholm är svårare. Routrarna står utspridda geografiskt i

Stockholm, tekniker måste åka fysiskt och starta om. . *(Vårt interna backupnät som ska kunna ge "console" terminal på distans fungerar inte, visar sig vara en bugg i vårt interna system.)*

16:30 Alla routers omstartade även i Stockholm. Nätproblemet är löst. Dock finns flera följdproblem, eftersom vi nu är flera på plats, löses följande parallellt:

1. De automatiska failoverlösningar som finns har gjort vad de kunnat för att få igång trafik, vilket gör att trafik kommer in på flera nya sätt & ställen. Orsakar problem för en del kunder från vissa ISP:er.
Detta löses genom att göra omroutningar i nätet så att trafik igen tar en mer optimal väg. Det skapar korta avbrott för användare när trafiken tar andra vägar genom internet. Det känns "ryckigt".

2. Det finns vissa rack i Falkenberg som står utan nätverk, visar sig vara feldragen kabel som gör att redundansen inte fungerade till dessa rack.

3. Internrouting till Cloudservers har problem, omstart av routing på en hel del noder krävs för att få fram nätverkstrafik till Cloudservers. Detta är ett följdproblem från det stora routerproblemet. Noderna fungerar som smårouters och har ärvt problemet från de större routrarna.

17:20 Vi ser nya problem i Stockholm, och tror att problemet är tillbaka. Visar sig dock bara vara överbelastning.

17:50 Nästan alla larm är lösta. (från 1033 st larm till 56 st i övervakningssystemet)

18:00-21:00 Vi fortsätter arbetet med att optimera inkommande trafik. I Falkenberg får vi tex. problem med BGP sessionen från Tele2 som är en stor leverantör hos oss vilket påverkar många kunder åter igen.
Vi väljer därför att helt strypa anslutningen mot Tele2, vilket stabiliserar nätet avsevärt.

Extrainkallad personal får åka hem och återgå på semester.

Vi upptäcker att supportsystemets mailserver blockerat sig själv pga för många fel. Vilket gör att inga nya tickets har skapats. När vi löser detta strömmar mail från oroliga kunder in.

21:00 - framåt

Ca 500 supportärenden finns registrerade, vi försöker svara på så många som möjligt och svara på de frågor som uppstått. En del kunder har följdproblem som vi försöker hjälpa till med.

En del Xen servers i cloudet rapporteras under kvällen nere. Det visar sig att i början av felet tappade inte alla noder nätverket på samma gång, vilket gjort att den automatiska failovern för Xen servers har flyttat (helt enl. plan) servers till de noder som fungerat. Dock har en del servers inte startats korrekt pga brist på resurser. Detta löses manuellt.

Slutsatser

Vi har under söndagen och måndagen haft möten och pratat igenom situationen. Det finns mycket vi kan bli bättre på, här kommer en kort genomgång av vad vi hittills identifierat.

Bevisligen är inte Stockholm och Falkenberg åtskilda tillräckligt. Detta är hösta prioritet på att lösa. Datacenter i Amsterdam och New York klarade sig utan problem eftersom de är mer åtskilda än Falkenberg & Stockolm. Ett fel som detta ska inte kunna, och får absolut inte sprida sig som det gjorde.

Nätverket är fortfarande komplext, eftersom vi har så många "siter". Ett mindre komplex nät hade gått snabbare att felsöka och lösa problem i. Detta problem har vi sett tidigare, och många åtgärder har gjorts för att göra det mindre komplext, men vi måste göra mer. Men vi vill samtidigt inte bli känsligare för DDOS attacker.

Det finns ett geografiskt problem i Stockholm som vi måste fundera igenom om vi kan göra annorlunda. VI måste även övervaka och testa våra backupvägar in till routrarna. Att inte dessa fungerade skapade nästan 1 timmes extra nertid för Stockholmskunder.

Vi kunde inte uppdatera driftinfo, inte heller fungerade supportmail. Även detta är katastrofalt. Vi kommer skapa en extern driftinformationssida. Vi har lärt oss att ärendesystemet kan sluta hämta nya ärenden, det har vi alltså redan nu bättre koll på för framtiden.

Vi höll visserligen uppdateringar på twitter, men det är inte fullgott. Uppdateringen via Twitter gick dock bra och vi känner att vi fick ut mycket viktig och riktig information den vägen. *(Man måste även förstå att samla ihop information från tekniker till korrekta statusuppdateringar tar viss tid. Teknikerna måste koncentrera sig på att lösa problemen i första hand.)*

Den tekniska beskrivningen på problemet är att vi troligtvis fått en läcka av externa routingregler in i vårt interna nät. *(Tänk hela internets alla routrar, laddas ner in i vårt interna nät).* Vi har inte lyckats återskapa problemet, men alla efteranalyser tyder på detta. Läckan kan bero på en bugg eller ett konfigurationsfel från vår sida, där man dock får anse att det sist nämnda är mest troligt. Vi rullade därför tillbaka våra routerkonfigurationer från backup i Lördag. Ett sådant fel ska dock inte få så stora konsekvenser som det fick, och det finns skydd att bygga mot detta som vi måste se över.

Det blir även väldigt hård belastning på våra nätverkstekniker i en sådan situation som uppstod. Vi bör utöka med ytterligare en person på detta område.

Till sist vill jag rikta ett stort tack till alla våra kunder som under eftermiddagen visade stort tålamod och kom med ryggdunk trots den kritiska situationen. Vi hade ett riktigt allvarligt avbrott och vi ser själva oerhört alvarligt på detta och vi jobbar redan intensivt med att ta så mycket lärdom som möjligt över lördagens händelser, samt sätter upp nya strategiska punkter för att motverka att liknande fel ska kunna drabba oss i framtiden.

Mvh

Glenn Johansson
vd, GleSYS Internet Services AB