

Hoge beschikbaarheid in de SURFconext authenticatie infrastructuur

Het gebruik van SURFconext introduceert een nieuwe 'schakel' in de communicatie tussen een instelling en een (cloud)dienst. Hierdoor ontstaat een extra afhankelijkheid voor het gebruik kunnen maken van diensten; immers, als de authenticatie via SURFconext niet werkt kunnen gebruikers niet op diensten inloggen. Hierom werkt SURFconext aan een zo hoog mogelijke beschikbaarheid van het platform, zodat de kans op uitval minimaal is. Echter is het SURFconext-platform slechts onderdeel van een grotere keten die uiteindelijk in zijn geheel moet werken om de gebruiker in te kunnen laten loggen.

Deze pagina legt uit hoe die keten precies werkt en wat per schakel kan worden gedaan om de beschikbaarheid zo hoog mogelijk te maken.

- Inleiding: sessies bij SURFconext
- Vijf verbindingen om rekening mee te houden
- Drie beschikbaarheidsscenario's
- Hoge beschikbaarheid
- Wat kan de instelling doen voor schakel 1 in de SURFconext federatie, de IdP
- Hoe zit het met de beschikbaarheid van schakel 2, (instellings-) diensten
- Een failover mogelijkheid realiseren voor schakel 3, SURFconext
- Samenvattend

Inleiding: sessies bij SURFconext

Als inleiding op beschikbaarheid is het van belang om eerst stil te staan bij het authenticatieproces van SURFconext en de bijbehorende sessies die actief worden na succesvolle authenticatie. [In het blog over de sessieduur van SURFconext](#) wordt dit als volgt beschreven:

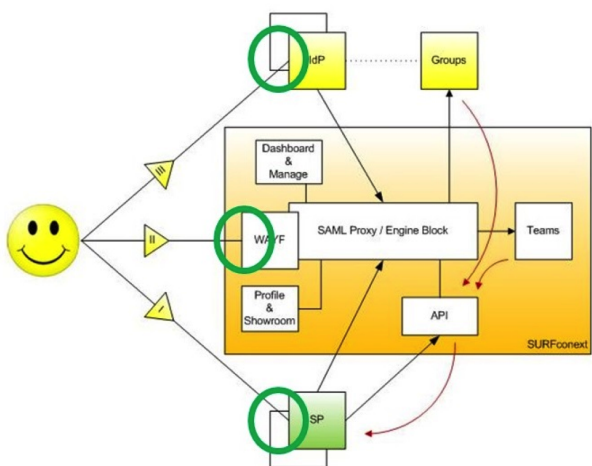
In de SURFconext authenticatie-infrastructuur is er niet sprake van één sessie, maar moet er rekening worden gehouden met drie verschillende sessies.

(I) Allereerst gaat een gebruiker (met zijn webbrowser) naar de dienst (SP). Op dat moment zal de webserver van de dienst de gebruiker doorsturen naar SURFconext.

(II) Aangekomen op SURFconext zal SURFconext de gebruiker doorsturen, eventueel via de WAYF, naar de login pagina van de instelling (IdP).

(III) Op de IdP, zal de gebruiker inloggen en na een succesvolle login wordt de sessie gestart. Via SURFconext wordt de gebruiker teruggestuurd naar de dienst. Op beiden wordt op die contactmomenten ook een sessie gestart.

De webbrowser heeft dus contact gehad met drie webserver, de SP, SURFconext en de IdP, die alle drie een eigen sessie zijn gestart, onafhankelijk van elkaar.



Vijf verbindingen om rekening mee te houden

Om het juiste inzicht in de mate van beschikbaarheid van diensten die op SURFconext zijn aangesloten te krijgen is het belangrijk om goed te begrijpen hoe de verbindingen tussen de verschillende partijen in een federatie verlopen. Uit bovenstaande authenticatieflow is duidelijk te zien dat er drie verbindingen essentieel zijn om een sessie te beginnen:

1. De verbinding tussen de webbrowser van de gebruiker en de dienst (SP)
2. De verbinding tussen de webbrowser van de gebruiker en SURFconext
3. De verbinding tussen de webbrowser van de gebruiker en de instellings-IdP

Er zijn nog meer pijlen getekend in het bovenstaande plaatje. Er is namelijk ook een pijl van de instelling naar de SURFconext SAML-proxy en van de SP naar de SURFconext SAML-proxy. Deze pijlen geven aan dat er een vertrouwensrelatie is tussen de IdP en SURFconext en tussen de SP en SURFconext. Die relatie is vastgelegd in de metadata die de systemen met elkaar uitwisselen. Deze relatie is echter out-of-band; de metadata hoeft niet real-time uitgewisseld te worden en kan heel goed eenmalig worden uitgewisseld en vanaf een lokale disk worden gelezen. Over deze verbinding zal dan ook geen data lopen op het moment dat een gebruiker in de authenticatieflow terecht komt.

Twee andere pijlen zijn wel weer van belang als het gaat om beschikbaarheid van diensten die op SURFconext zijn aangesloten. Een dienst kan namelijk het groepslidmaatschap van de gebruiker opvragen aan de SURFconext VOOT-omgeving. Dit groepslidmaatschap kan bestaan uit lidmaatschap van SURFconext Teams en uit lidmaatschap van instellingsgroepen. Voor dat laatste bevraagt SURFconext het groepencomponent van een instelling. Ingeval een dienst van groepen gebruik maakt zijn er dus twee additionele verbindingen om rekening mee te houden:

4. De verbinding tussen SURFconext en de groepprovider van de instelling
5. De verbinding tussen de dienst en de VOOT-omgeving van SURFconext

Drie beschikbaarheidsscenario's

Het **authenticatiescenario**, het moment dat een gebruiker inlogt op een dienst en via SURFconext wordt doorgestuurd naar zijn instelling, is het meest gebruikte scenario waarbij SURFconext een cruciale rol speelt. De gebruiker komt in dit geval langs drie verbindingen die alle drie beschikbaar moeten zijn. De zwakste schakel in deze drie verbindingen bepaalt de mate van beschikbaarheid. Dit scenario herhaalt zich iedere keer als een gebruiker op een dienst inlogt; de drie verbindingen worden dus altijd gebruikt. Ook in geval van Single Sign-On, als wordt herkend dat de gebruiker al eerder ingelogd is geweest, dan wordt hij direct weer doorgestuurd, zonder dat hij daarbij een extra actie hoeft te ondernemen (bijv. het opnieuw invullen van zijn gebruikersnaam en wachtwoord). Maar zelfs dan worden de drie verbindingen gebruikt.

Als een gebruiker eenmaal is ingelogd op een dienst zal hij gebruik maken van die dienst. Zolang de sessie actief blijft hoeft de gebruiker niet opnieuw contact te hebben met SURFconext of zijn instelling. In dit scenario, het **gebruiksscenario**, is dan ook alleen de verbinding tussen de webbrowser van de gebruiker en de SP van belang. Deze verbinding bepaalt de mate van beschikbaarheid (van de SP).

Als een dienst gebruik maakt van groepen, het **groepenscenario**, dan is er sprake van een extra verbinding tussen de SP en SURFconext. Als de informatie over groepslidmaatschappen direct vanuit de instelling komt, is er sprake van nog een extra verbinding, namelijk tussen SURFconext en de groepprovider van de instelling. Een dienst die gebruik maakt van groepen, zal typisch eenmalig SURFconext om groepslidmaatschap vragen om op basis daarvan autorisatie te kunnen doen. Deze verbinding zal meestal worden aangeroepen vlak nadat de gebruiker geauthentiseerd is; daarna zal de verbinding typisch niet veelvuldig meer gebruikt worden.

Het zwaartepunt van deze beschikbaarheidsscenario's ligt bij het gebruik van de dienst, het gebruiksscenario. Dit komt omdat een gebruiker ieder moment van de dag in een dienst aan het werken kan zijn, bijvoorbeeld om een Wiki of leeromgeving bij te werken of om bestanden van en naar een externe opslagdienst te verplaatsen. De SP dient op al deze momenten van de dag beschikbaar zijn.

Het authenticatie- en groepenscenario zijn ondersteunend hieraan. Deze zullen typisch aan het begin van de dag (en begin van de middag) veel gebruikt worden, op het moment dat een gebruiker aan zijn werk begint en een dienst opstart. Hoe vaak een gebruiker opnieuw moet inloggen en hoe belangrijk de ondersteunende scenario's hierbij dus zijn, is afhankelijk van de geldigheidsduur van de sessie, en of de gebruiker bijvoorbeeld tussendoor zijn browser afsluit. [Lees het blog over sessieduur voor meer informatie hierover.](#)

Hoge beschikbaarheid

Wat is eigenlijk hoge beschikbaarheid?

In het algemeen zorgt een hoge beschikbaarheid van een dienst ervoor dat een defect component in een keten geen invloed heeft op de dienstverlening en functionaliteit van de dienst omdat een ander component deze automatisch vervangt. De beschikbaarheid van een dienst is zo sterk als het zwakste component.

Op het hoogste niveau zijn de drie genoemde schakels betrokken in de SURFconext-federatie (IdP, SURFconext, SP). Daarnaast zijn er meerdere lagen waarop beschikbaarheid betrekking heeft, ook binnen deze lagen bepaalt de zwakste schakel de sterkte. De lagen die te onderscheiden zijn, zijn: de locatie, het netwerk, de servers, de applicaties en de data. Typische beschikbaarheidsvragen die vervolgens over deze lagen gesteld kunnen worden, zijn:

1. Locatie (dit betreft de beschikbaarheid van het datacentrum). Als vanwege een stroomstoring of iets soortgelijks een datacentrum niet beschikbaar is, is er dan een tweede datacentrum waar naar uitgeweken kan worden?

2. Netwerk. Als de connectiviteit van een netwerk naar de servers van een dienst onderbroken wordt, is er dan een alternatief netwerk voor deze servers?
3. Servers. Als een onderdeel van een server uitvalt (bijvoorbeeld een harddisk), is er dan een vervangend onderdeel? Of als de server helemaal uitvalt, is er dan een vervangende server?
4. Applicaties. Als een applicatie dienst weigert (vanwege een tekort aan geheugen bijvoorbeeld), is er dan een vervangende applicatie, of een vervangende server beschikbaar?
5. Data. Is de data die de applicaties gebruiken nog steeds beschikbaar als de database wegvalt?

Hoge beschikbaarheid van een laag is te bereiken door:

1. Een cluster te maken. In een cluster zijn er meerdere componenten die dezelfde functionaliteit leveren. De dienst wordt geleverd door het cluster als geheel. Als een component binnen het cluster uitvalt, dan neemt een ander component in het cluster de dienst over, zonder dat het geheel daar last van heeft.
2. Een loadbalancer in te zetten. De loadbalancer (of reverse proxy) krijgt een verzoek binnen en verdeelt deze over achterliggende componenten. Als een van deze achterliggende componenten uitvalt, dan zal de loadbalancer deze niet meer gebruiken en zullen de andere componenten de load van dit component overnemen.
3. Een failover in te zetten. Als een component uitvalt dan wordt dat gedetecteerd door het failover component, die vervolgens de functionaliteit overneemt, totdat het oorspronkelijk component weer actief wordt. Failover kan op systeemniveau worden gerealiseerd door een zogenaamde heartbeat tussen het actieve en het failover component op te zetten. Als de heartbeat verdwijnt, dan dient de failover in actie te komen. Failover kan ook met virtualisatietools, zoals VMware of Hyper-V worden gerealiseerd.

Meestal zal een combinatie van technieken nodig zijn om hoge beschikbaarheid te garanderen. Als bijvoorbeeld een loadbalancer de verzoeken aan meerdere achterliggende servers doorspeelt, dan is er vaak ook een failover loadbalancer nodig die de loadbalancing taak overneemt in het geval dat daar een storing optreedt.

Uiteindelijk is het verhogen van de beschikbaarheid een afweging tussen de waarschijnlijkheid dat een component uitvalt, de mogelijke tijdsparren van de uitval en de kosten die ermee gemoeid zijn om een verhoogde beschikbaarheid te bereiken. In [een onlangs verschenen blog over de betrouwbaarheid van SURFconext](#) wordt uitgelegd hoe bij SURFconext deze technieken ingezet worden.

Wat kan de instelling doen voor schakel 1 in de SURFconext federatie, de IdP

Aan de hand van een aantal vragen die hieronder over beschikbaarheid gesteld worden kan een instelling nagaan of hoge beschikbaarheid in combinatie met SURFconext wenselijk, dan wel gerealiseerd is.

Voor de instelling is het van belang dat de IdP van de instelling altijd bereikbaar is. Dit is de centrale dienst die de loginpagina voor de gebruikers verzorgt en de gebruiker met SAML authenticiseert.

Om een hoge beschikbaarheid voor een IdP te realiseren is het belangrijk om allereerst te kijken naar de locatie en het netwerk. Daarvoor kan de volgende vraag gesteld worden: is er voor iedere gebruiker, op de campus en thuis, een redundante netwerkvoorziening? En is er een uitwijkmogelijkheid naar een andere locatie? (Of treden er bij een onbereikbare locatie andere protocollen in werking?) Deze vraag is niet specifiek op de IdP van toepassing, maar op de hele ICT-infrastructuur van de instelling en zal daarom vaak op een hoger niveau al beantwoord kunnen worden.

Vervolgens kan gekeken worden naar de servers en applicaties. De applicatie die de IdP dienstverlening levert is vaak een product als Microsoft ADFS, SimpleSAMLphp of Shibboleth IdP. De volgende vraag is: is deze applicatie op meerdere systemen actief? Dat kan in loadbalancing-modus zijn, waarbij er twee of meer instanties van dezelfde applicatie altijd actief zijn. Of dat kan in failover modus zijn, waarbij er één instantie actief is en één stand-by. Daarnaast: als er een loadbalancer wordt gebruikt, is daar dan ook een failover van?

Een IdP maakt vaak gebruik van een achterliggende database met gegevens van gebruikers, zoals een LDAP-server of een Active Directory. Is deze database op meerdere servers actief? Dat kan, net als bij de IdP, in loadbalancing of failover modus zijn. Voor databases geldt vaak dat ze ook in cluster mode kunnen opereren. Active Directory is bijvoorbeeld in zichzelf al ingericht op hoge beschikbaarheid. Door het opzetten van een tweede domain controller (DC) wordt de directory al gerepliceerd. Met behulp van DNS-loadbalancing kunnen beide DC's worden gebruikt. Is er een tweede domain controller (of ander cluster) actief?

Wat voor de IdP geldt, geldt ook voor de instellingsgroepdienst. Deze groepdienst is erg vergelijkbaar met de directory dienst. Deze dienst biedt groepsinformatie aan, middels het VOOT-protocol (in aanvulling op een LDAP die gebruikersinformatie levert). Het kan een aparte server zijn die deze groepdienst aanbiedt of de dienst kan geïntegreerd zijn in de directory. De vraag die gesteld moet worden is: heeft de groepdienst ook een failover?

Hoe zit het met de beschikbaarheid van schakel 2, (instellings-) diensten

Voor de beschikbaarheid van een dienst die uit de cloud wordt afgenomen, is de leverancier verantwoordelijk. Hier heeft instelling weinig grip op. Daarom maakt een instelling hier vooraf duidelijke afspraken over met de leverancier. Voor een deel van de diensten die via SURFconext wordt afgenomen, geldt dat SURFmarket namens de instellingen dit soort afspraken contractueel vastlegt. Deze afspraken staan in het [juridisch normenkader cloudservices](#). Hierin staat bijvoorbeeld dat de leverancier verantwoordelijk gesteld wordt voor de gewenste beschikbaarheid van de gegevens en dat van de leverancier wordt verwacht dat er passende maatregelen worden genomen om hoge beschikbaarheid te garanderen.

Diensten die instellingen zelf aan SURFconext koppelen kunnen voorbereid worden op hoge beschikbaarheid. Zoals eerder genoemd is dit ook erg belangrijk omdat het zwaartepunt van de beschikbaarheidsscenario's bij de dienst zelf ligt.

Ook hier geldt dat alle lagen: locatie, netwerk, servers, applicaties en data afzonderlijk bekeken moeten worden. Waarbij aangetekend moet worden dat voor een SP, misschien nog wel meer dan voor een IdP, geldt dat deze voorbereid moet zijn op variaties in belasting van het systeem. Daarom zal een SP eerder kiezen om de servers in een loadbalanced opstelling neer te zetten dan in een failover setting.

Een failover mogelijkheid realiseren voor schakel 3, SURFconext

Ervan uitgaande dat de instelling de IdP en SP voor hoge beschikbaarheid heeft opgetuigd, dan is er vervolgens, vanuit de instelling bekeken, nog een failover component in te zetten voor SURFconext. Bij gebruik van een instellingdienst beheert de instelling namelijk zowel de SP als de IdP, maar niet SURFconext. Wat als het onverhoopt toch gebeurt dat het SURFconext component uitvalt? Dan breekt de keten voor de instelling.

In het geval van uitval van SURFconext zal de SP moeten detecteren dat SURFconext niet beschikbaar is. Op dat moment dient de SP zichzelf te her-configureren, zodanig dat bij een nieuwe login de gebruiker niet via SURFconext, maar rechtstreeks op de IdP van de instelling kan inloggen. Daarvoor moet er initieel ook een trust relatie opgezet worden tussen de SP en IdP van de instelling. Zij moeten metadata uitgewisseld hebben zodat een SAML-request direct doorgestuurd kan worden.

Merk op dat deze failover alleen mogelijk is voor instellingsdiensten en gebruikers van de eigen IdP. Gasten zullen door de failover ook naar de IdP van de instelling worden doorgestuurd, maar kennen daar geen loginmogelijkheid en komen dus eigenlijk op de verkeerde plek. Daarnaast is het voor de SP moeilijk te detecteren of SURFconext wel of niet beschikbaar is voor de gebruiker omdat de gebruiker een andere route naar SURFconext heeft dan de dienst zelf. Het is dus, net als altijd, noodzakelijk de kosten van deze aanpassing af te wegen tegen de opbrengsten en de waarschijnlijkheid van deze gebeurtenis.

Samenvattend

Hoge beschikbaarheid in de SURFconext authenticatie infrastructuur is te realiseren door te focussen op de onderliggende componenten. Daarbij zijn er meerdere lagen waarin rekening gehouden moet worden met hoge beschikbaarheid. Er zijn verschillende technologieën om de beschikbaarheid per laag te verhogen. De instelling kan hier zelf een rol in spelen op IdP-niveau en voor instellingsdiensten. SP's zijn zelf verantwoordelijk voor de beschikbaarheid van hun dienst en SURFconext draagt zijn steentje bij door constant aan hoge beschikbaarheid en betrouwbaarheid te werken.

Verder lezen:

- [Meer over sessiegebruik in SURFconext](#)
- [Meer over betrouwbaarheid van SURFconext](#)