Preliminary findings and recommendations from the Token Trust and Traceability Working Group

Matt Doidge[1,*], *David* Crooks[2], *David* Kelsey[2], *Maarten* Litmaath[3], *Linda* Cornwall[2], *Mischa* Salle[4], *Marcus* Hardt[5], and *Tom* Dack[2]

[1]Lancaster University
[2]STFC
[3]CERN
[4]NIKHEF
[5]KIT

Abstract.
Created in 2023, the Token Trust and Traceability Working Group (TTT) was formed in order to answer questions of policy and best practice with the ongoing move from X.509 and VOMS proxy certificates to token-based solutions as the primary authorisation and authentication method in distributed computing environments. With a remit to act in an investigatory and advisory capacity alongside other working groups in the token space, the TTT is composed of a broad variety of stakeholders to provide a breadth of experience and viewpoints. While the requirements of grid sites, users, identity providers and virtual organisations to be able to trace workflows have remained largely the same in a token paradigm as to one using X.509 certificates, tokens provide a new set of challenges, requiring a rethink and restructure of the policies and processes that were defined with just X.509 and VOMS in mind, in order to meet these requirements in the new context. After providing an overview of the current status of the token trust landscape we will detail the initial findings, future plans and recommendations to be made by the TTT. This will include best practice for sites and identity providers, suggestions for token development, and methodologies for tracing token usage by system administrators within common grid middleware stacks.

1 Introduction

The Token Trust and Traceability Working Group (TTT) was formed to fulfil a growing need in distributed computing communities, such as WLCG, EGI and IGWN[1], to tackle questions arising from the move to token based AAI (Authentication and Authorisation Infrastructure). The previous work in the area, conducted by the Traceability and Isolation WG [1], had focussed on the previous AAI management paradigm of VOMS[2] and X.509 certificates. The aim of the TTT is to adapt and expand the work of the previous WG within the new AAI paradigm of tokens.

[*] e-mail: m.doidge@lancaster.ac.uk
[1] International Gravitational Wave Observatory Network
[2] Virtual Organisation Management Service

2 Tokens, Trust and Traceability

2.1 Token Background

Introduced by the commercial sector (Google, Microsoft), token based solutions are being adopted by many newly designed distributed computing infrastructures, such as SKA[3] or NFDI[4]. Also existing ones, such as EuroHPC, and, as described by the WLCG AuthZ working group [2][3], WLCG, are set to transition from X.509 based authentication and authorisation to one using Tokens. The activities are centered around OpenID Connect Providers (OPs), such as Indigo IAM [4], RCIAM [5], the GEANT Core AAI Platform [6], Unity [7], Perun [8], Keycloak [9] and CILogon [10]. During this time many distributed Computing Infrastructures and Research Communities were moving to, or had already implemented token based authentication and authorization. Within the wider landscape, tokens have been used for many years across industry, and are a well established technology.

Tokens within our context are described in more detail elsewhere [11]: when referring to tokens in this document we refer either to an "Access Token", usually a JSON Web Token (JWT) comprised of an encoded JSON with a predefined structure, or a "Refresh Token", usually an opaque string. The former is used to grant access to the bearer of the token on behalf of and authorized by a user, while the latter can be used by the specific client the token was issued to, to request new associated Access Tokens from the OP or OAuth2 Authorization Server (AS).

The structure of an Access Token matches a predetermined profile. The token is provided to an OAuth2 client, usually on behalf of a user and signed by an "Issuer" - the service that provides tokens, i.e. the OP or AS. Within WLCG the Issuer is unique per VO, but this is generally not true for other communities. The structure of a JWT for use as an OAuth2 bearer token is suggested by RFC9068 [12]. It contains fields such as (not all from RFC9068):

- version: The profile version.
- scope: The action(s) that the token is able to authorise.
- aud: The audience claim, indicating on which service(s) the token is authorised to perform the action(s).
- iss: The issuer of the token.
- jti: A unique identifier for the token.
- sub: token subject, equivalent of the Distinguished Name (DN) in certificates.
- Time and Date information describing the period the token is valid for: e.g. iat "Issued At", nbf "Not Before", exp "Not After".
- client_id: indicating the client the token was minted for.
- Other fields dependent on the Profile.

In addition Tokens can be broadly described as "Offline Introspectable" - the information within them can be extracted by a client service, or "Opaque" - requiring a client service to contact the Issuer to verify a token.

2.2 Token Trust

The introduction of Token based authentication brings changes not only in the technology used by clients and AAI provider, but also in the underlying methodology of how authorisation decisions can be made, in particular when compared to X.509 Proxy technologies.

[3] Square Kilometre Array
[4] Nationale Forschungsdateninfrastruktur

With token attributes such as Scopes and Audiences or Entitlements there is greater capacity for granularity on authorisation decisions, making it possible to describe in greater detail the capabilities being granted. The Issuer has a more central role, with a greater possibility of being called back to mid-workflow.

With new technology come new risks arising from inexperience, and a desire to not squander the potential benefits provided from the shift to token based AAI.

For example, unlike certificates, tokens were not originally designed to be verified offline but to have introspection ([13]) used instead, and mechanisms such as certificate revocation lists are not known in the token world.

2.3 Traceability Requirements

In the final findings of the WLCG Traceability and Isolation WG an emphasis was placed on the importance of, and thus the need for, full traceability - for any action on a service there is a requirement to be able to deduce what action was performed, when, and by whom. This requirement fundamentally has not changed with the transition to tokens. However, the methods and flows that one would follow to trace an activity will have. One aim of the TTT is to document these changes in activity tracing, and ensure that clients and services are logging and preserving the needed information.

3 Risk analysis, per workflow approach

Through discussion within the group, and following on from previous examples within the area [14], it was decided that the optimal approach for formulating recommendations was through adopting a Risk Analysis based methodology. Whilst still to be formalised, the proposed approach would consist of evaluating recommendations in terms of associated risks, in terms of operation and practical issues as well as more obvious risk factors such as chance of credential exposure or misuse.

As even within a single community or user group the token use cases can vary widely, it was deemed that this risk analysis approach would be performed "per workflow" rather than attempt to apply it to any group as a whole. Examples of workflows include bulk data transfer, data access by individuals, remote job submission, and data access by jobs.

3.1 Token Attributes as Variables

As described, a typical JWT token provides multiple attributes. Some of these attributes can be considered as to be "tunable" variables, evaluated in such a way that they can be considered either more, or less, secure. This allows for visualisation of a token use case's "trustability".

3.2 Visualising Token Trust Parameters

In Figure 1 we take three of the token attributes that can most simply be evaluated and adjusted and, for the purposes of this model, consider them to be orthogonal. With each parameter it can be considered that an increase in the magnitude - representing an increase in token lifetime, a widening of token scope or a loosening of the token audience - is representative of an increase in the risk of using a token with those parameters applied. Within such a model the desirable outcome to minimise risk involves choosing a point as close to the "origin" as possible.

This model can be used to roughly judge the risk of an access method and its token attributes. Applying the logic of this model to the long lived X.509 proxies commonly used

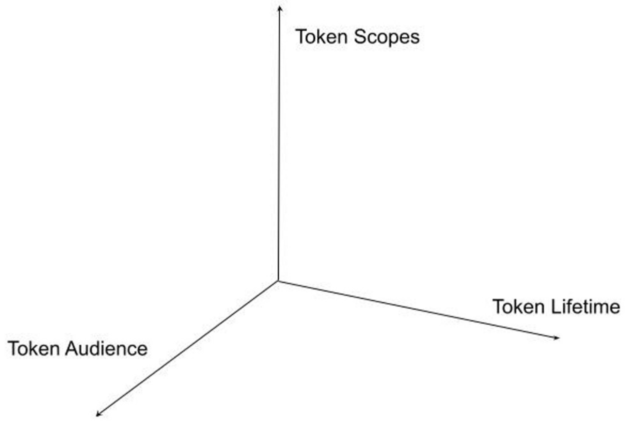

Figure 1. Visualisation of the parameter space of Token Lifetime, Scopes and Audience. Parameter values closer to the origin are considered safer, i.e. having a lower risk.

within grid jobs where the scope is large, there is effectively no defined audience and the lifetime is measured in days, this mode of operation is considered rather high risk. Caution should be taken when directly comparing X.509 and Token workflows, as not all factors are "like-for-like".

For planning purposes using the logic of this model, an operational need to loosen requirements on one axis could be compensated for by tightening on the other two axes. For example, a need for a longer token lifetime can be offset with strict audience or very limited scope in order to make the remaining risk acceptable. This model should be applied alongside other considerations when analysing the risk of a workflow, such as capacity for token revocation or level of token exposure.

3.3 Token Lifetime

Token Lifetime is the most easily quantifiable of the token attributes, and one that requires extra consideration due to the potential for operational impacts. These impacts are also being looked at by other groups [15].

The operational concerns involving Token Lifetime arise from the increased requirements for token renewal when using shorter lifetime tokens in long lasting workflows. This increases the potential load on the Token Issuer by increasing the frequency of requests, as well as the chances of failure when the Issuer does not respond at a critical time.

The common consensus is to try and avoid overly long lasting tokens (longer than 6 hours) in particular if there is no means of revocation, and for various use cases this is sufficient. However, data-intensive computing has at least two use cases, common across groups, where a longer lifetime is desired: the access tokens required for bulk data transfers, and the tokens used by computational jobs for reading input data and writing job output to storage. These processes could both be spread over a period of multiple days after initiation, over a time period that might be difficult to precisely predict. There are multiple solutions that attempt to avoid or mitigate the use of straight very long lasting tokens, which are all candidates for evaluation through a token risk analysis framework.

3.4 Other requirements and considerations

There are considerations that cannot easily be parametrised that still need to be considered within the risk models. For example Token Exposure - how exposed a token is, through storage location and transfer, to being intercepted or accessed, is one such consideration.

https://doi.org/10.1051/epjconf/202533701080

Workflows that require tokens to be stored on external, unmanaged services may require stricter considerations for, say, their scopes and lifetimes, unless secured in additional ways.

Mechanisms for Token Revocation or User Ban lists, particularly as a mitigation for long life tokens, are an area to be explored. No mechanism exists yet within a grid environment, but the TTT is investigating the plausibility and technical details.

The roles, responsibilities and Incident Response capabilities of the Token Issuers are another area that doesn't neatly fit into the risk model, but which is vitally important. The Token Issuer is at the center of all token flows, and any tracing of an activity would require interacting with those running the Issuer service. This becomes especially true in the event of a Cybersecurity Incident. The response expectations for the Issuer need to be defined, as well as guidance on the release of information from the Issuer, what can be shared and whom it may be shared with.

A consequence of using tokens that are a more precise fit for their intended usage is a potentially substantial increase in the amount of Tokens found in an infrastructure. The optimum between large amounts of highly specialised tokens (e.g. one per file) versus a low amount of tokens with a broad range of capabilities (e.g. X.509 Proxies as the extreme case) still needs to be determined, taking into account for example the consequences on traceability.

A broader area for the TTT is documentation - Tokens are still considered a new technology within the Grid space. Increased and improved documentation is required in order to build trust in the new way of working, as well as reduce the chances of human error through lack of knowledge.

4 TTT WG Future Plans

The Token Trust and Traceability WG is still within its early days. Our goals for 2025 and beyond will involve a number of steps, such as:

- Formalise and mature the risk model.
- Maintain involvement with other working groups and efforts in the area of AAI.
- Create and encourage documentation.
- Continue to meet and debate the ongoing changes in the token space [16].
- Formulate requirements on traceability of user activities across infrastructures.

The ultimate aim of the TTT over the coming years is to produce tangible recommendations drawn from strong logical foundations and enable the informed creation of policies on the use of Tokens within our diverse research computing environments.

References

[1] Brillault, Vincent et al, “WLCG Traceability and Isolation WG: recommendations”, https://indico.cern.ch/event/739880/contributions/3470862/attachments/1876966/3093414/WLCG_Traceability_and_Isolation_WG__recommendations.pdf (Accessed Feb 2025)

[2] Ceccanti, Andrea, Vianello, Enrico Caberletti, Marco and Giacomini, Francesco, “Beyond X.509: token-based authentication and authorization for HEP”, **EPJ Web Conf. Volume 214** (CHEP 2018) https://doi.org/10.1051/epjconf/201921409002

[3] Brian Bockelman, Andrea Ceccanti, Ian Collier, Linda Cornwall, Thomas Dack, Jaroslav Guenther, Mario Lassnig, Maarten Litmaath, Paul Millar, Mischa Sallé, Hannah Short, Jeny Teheran and Romain Wartel, “WLCG Authorisation from X.509 to Tokens”, **EPJ Web Conf. Volume 245** (CHEP 2019) https://doi.org/10.1051/epjconf/202024503001

[4] "INDIGO Identity and Access Management Service (IAM)": https://indigo-iam.github.io (Accessed March 2025)

[5] RCIAM: https://github.com/rciam (Accessed March 2025)

[6] GEANT Core AAI Platform: https://github.com/geant-core-aai (Accessed March 2025)

[7] Unity IDM: https://unity-idm.eu (Accessed March 2025)

[8] Perun AAI Platform: https://perun-aai.org (Accessed March 2025)

[9] Keycloak: https://www.keycloak.org (Accessed March 2025)

[10] CILogon: https://www.cilogon.org/ (Accessed March 2025)

[11] WLCG Common JWT Profiles: https://zenodo.org/records/3460258 (Accessed March 2025)

[12] RFC9068: https://www.rfc-editor.org/rfc/rfc9068.html (Accessed March 2025)

[13] RFC7662: https://www.rfc-editor.org/rfc/rfc7662.html (Accessed March 2025)

[14] Cornwall, Linda, "D4.4 Security Risk Assessment of the EGI Infrastructure" https://documents.egi.eu/public/ShowDocument?docid=863 (Accessed March 2025)

[15] AARC G081 Document: Recommendations for Token Lifetimes https://docs.google.com/document/d/1U9vvJfWuE8oO7u0FcGVGr3KySvBqwjnkzKO8TKzgoX4 (Accessed March 2025).

[16] Token Trust and Traceability WG: https://github.com/TTT-WG/TTT-WG (Accessed March 2025)