Scalability and Performance

Scalability

What options does Alma provide regarding scalability and performance?

Alma is designed for scalability; because it is deployed in the cloud environment, there are no system limitations as to the number of libraries or their capacity. This architecture allows for:

- Alma to be deployed in multiple instances in each data center, without limitation on the number of parallel instances that may be deployed; and
- Each Alma instance can contain many libraries (depending on their size) in a multi-tenant architecture.

This allows us to scale Alma in multiple levels:

- An Alma instance can scale:
  - Horizontally – by adding additional application servers and database servers; and
  - Vertically - by adding additional cores, RAM, etc., to existing servers.
- An additional Alma instance can be deployed as needed via automatic deployment tools.

When a new institution is added to Alma, it is deployed in the most appropriate instance, based on its volume, and new instances are opened in advance as needed, based on our pipeline.

The Alma SaaS environment is built to handle usage fluctuations and peaks in several ways:

- Access to the Alma application is done via a load balancer that routes customers to an available application server;
- Online transactions, batch jobs and reports are each performed via dedicated resources to prevent disruption to online transactions caused by heavy jobs or reports run by users; and
- Deployment of end-point monitoring tools for high visibility.

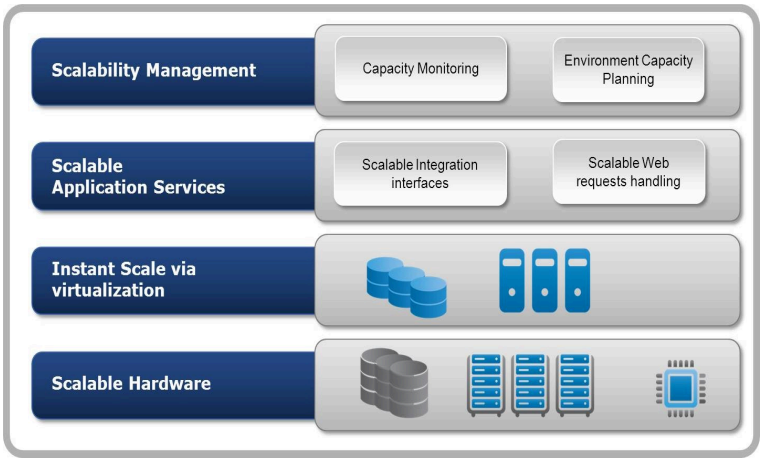
Our monitoring tools help us gauge performance, and in the event of any service degradation, our cloud engineers are supplied with the appropriate tools. For example, if a server's resources are being consumed at a sudden high rate, an engineer is notified, and can change the load balancer rules to allocate more resources for online transactions.

While Alma and Primo are sized for the hosted environment according to user parameters provided by the institution, the application can handle fluctuations as described above.

Ex Libris is committed to providing our customers with a highly secure and reliable environment for our cloud-based solutions. We have developed and deployed a multi-tiered security model that covers all aspects of cloud-based Ex Libris systems. The security model and controls are based on international protocols and standards as well as industry best practices, such as ISO/IEC 27001, the standard for information security management systems (ISMS). All communication between browsers and the Alma cloud is secured. Personal patron information (such as ID, email, address) is kept securely encrypted in the Alma database.

How does Alma manage capacity at an infrastructure level?

Alma is based on the multi-tenancy architecture in which one instance of Alma supports multiple institutions. To allow for growth in a cost-effective manner, Alma is designed to scale throughout its layers:



Scalability management:

Alma's cloud engineers oversee capacity planning and system monitoring. To plan for growth, the cloud engineers track our existing and ongoing implementations, as well as our future pipeline, to make sure we have the required capacity. At the same time, they monitor our existing customers' growth and the system's performance trends. This approach ensures that Ex Libris can provision more hardware to meet our customers' needs in real time.

Alma is designed for scalability:

Because Alma deployed in the cloud environment, there are no system limitations as to the number of customers or their capacity. The architecture allows for:

1. Alma to be deployed in multiple instances in each data center, without limitation on the number of parallel instances that may be deployed; and
2. Each Alma instance can contain many libraries (depending on their size) in a multi-tenant architecture.

This allows us to scale Alma in multiple levels:

1. An Alma instance can scale:

Horizontally – by adding additional application servers and database servers; and Vertically - by adding additional cores, RAM, etc., to existing servers.

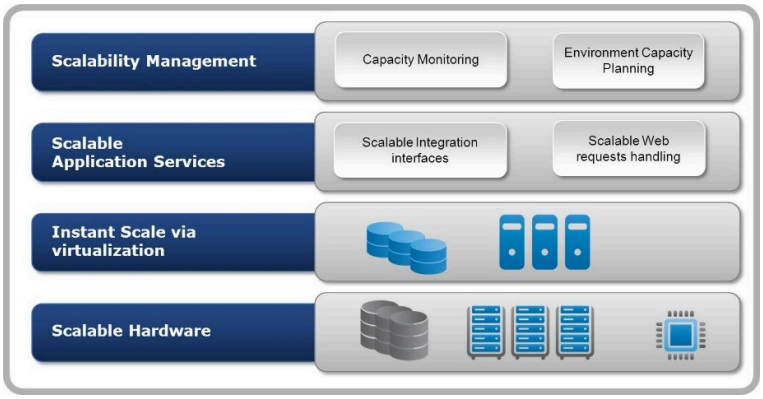
2. An additional Alma instance can be deployed as needed via automatic deployment tools.

When a new institution is added to Alma, it is deployed in the most appropriate instance, based on its volume, and new instances are opened in advance as needed, based on our pipeline.

Alma's multi-tenancy architecture ensures that Alma can expand to meet the needs of a growing organization.

What provision is provided for increases in users served and collections managed?

To allow for growth in a cost-effective manner, Alma is designed to scale throughout its layers:



Alma's multi-tenancy architecture ensures that Alma can expand to meet the needs of a growing organization or a growing collection, and Ex Libris' cloud management practices for capacity monitoring and request handling will support the variances experienced by the institution over the academic year.

If records are held permanently how are issues of increasing capacity handled?

Alma as a cloud-hosted multitenant solution, increasing capacity is not an issue. Alma's cloud engineers oversee capacity planning and system monitoring. To plan for growth, the cloud engineers track our existing and ongoing implementations, as well as our future pipeline, to make sure we have the required capacity. At the same time, they monitor our existing customers' growth and the system's performance trends. This approach ensures that Ex Libris can provision more hardware to meet our customers' needs in real time. Alma is designed for scalability.

Performance

What is the recommended bandwidth?

The Ex Libris data centre utilizes 1G bandwidth as its backbone, and works with multiple ISP vendors (up to 8) at every point in time. As with every SaaS vendor, we measure the performance of our servers; we do not, however, have control of the 'last mile' at the institution level. We can report that so far our customers report a high performance level for all operations that met web application expectations. We also monitor our application and server throughput 24/7 in order to make sure that transactions are handled and sent to the user browser as expected.

How does Alma address performance issues when there is a high workload?

Because a multi-tenancy environment serves many customers, we designed Alma to provide flexibility.

This is done in several ways, a few examples of which are:

- The data is saved in the Character Large Object (CLOB) data type, which allows for an unlimited number of data elements. An example of such usage is all MARC records as well as inventory objects;
- Many of the entities in Alma have a tab for notes, allowing for an unlimited number of notes records.
- Whenever required, values for fields are managed in a mapping table to allow for full customization on the institution level.

For regular fields, each has its data types as any typical application would have. In most of the cases, changes in these

types of fields will entail some level of development per specific case, but the Alma architecture allows for such changes without impact on the production environment.

How is performance monitored?

Ex Libris monitors the performance of Alma 24X7X365. The monitoring of Alma covers both the infrastructure and business transactions level. That is, our 24X7 hub monitors not only the connection performance to Alma and our servers' side response performance but also the performance of business transactions such as z39.50 connection to Alma, circulation transactions performance and more.

As with every multi-tenant solution, Alma has Governance Thresholds that ensure that no single institution negatively impacts other Alma institutions, prevent performance degradation and help reduce the risks of malicious attacks.

Such Governance Threshold exist in a few areas of the system for example importing and exporting data to/from Alma:

	Peak (working day) - records per hour	Off-Peak (night) – records per hour
Export	0.3M	1.2M
Manipulate	0.15M	0.6M
Import (**match by system ID**)	0.15M	0.6M

How does Alma manage peaks and spikes in workload over varying periods of time, including seconds, minutes and hours?

The Alma SaaS environment is built to handle usage fluctuations and peaks in several ways:

- Access to the Alma application is done via a load balancer that routes customers to an available application server;
- Online transactions, batch jobs and reports are each performed via dedicated resources to prevent disruption to online transactions caused by heavy jobs or reports run by users; and
- Deployment of end-point monitoring tools for high visibility.

Our monitoring tools help us gauge performance, and in the event of any service degradation, our cloud engineers are supplied with the appropriate tools. For example, if a server's resources are being consumed at a sudden high rate, an engineer is notified, and can change the load balancer rules to allocate more resources for online transactions.

Can Alma commit to 99.5% availability?

As part of our SLA we provide commitment of annual Uptime Percentage of at least 99.5%.

Uptime Percentage - means Uptime expressed as a percentage, calculated in accordance with the following formula:

Uptime Percentage = X /(Y–Z) × 100

Where:

X = Uptime

Y = number of minutes in a year

Z = The duration (in minutes) of any SLA exclusions during the year (such as planned maintenance window)

In practice we have constantly exceeded the annual 99.5% uptime SLA since the first day of going live.

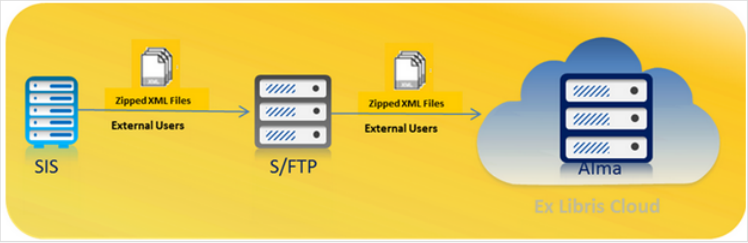
Batch Loading

How does Alma enable simultaneous batch operations across multiple institutions?

Batch jobs and processes in Alma are designed for multi-tenancy and distributed to ensure there is no impact on other institutions' processes. Dedicated batch-jobs servers and resources are allocated for batch job processes, thus eliminating the risk of interference with institutions' processes or resources. This architecture means that Alma can scale virtually infinitely, as the only barrier is server allocation, and there is no dependency on the institution's processes environment.

How can data be batch loaded?

The loading of user data from an external system(s) is performed using zipped XML files that are placed at a predefined, secure FTP location. Alma fetches the files, parses them, and updates external users according to the input file and the parameters defined in the external system profile(s). The diagram below illustrates the communication between the external system(s) [SIS] and Alma. The loading of external users into Alma can be performed in one of two modes: (i) import, or (ii) synchronize. The import mode is a one-time load, used to initially create new external users. It is intended to be used only when you have a file of users you know are new, such as during the migration process, when you want to load users from your legacy system into Alma. The synchronize mode is an on-going load, used to update external users and add new ones.



Batch Loading in other areas

Batch loading is mainly used in the following areas:

- Acquisitions Embedded Order Data (EOD) files – Alma enables defining Import Profiles that define all aspects of the EOD import process, including:
- Import process definitions (FTP location, scheduling, normalization, validation, matching)
- Acquisitions records mapping definitions – which fields in the EOD files map into which PO Line fields
- Bibliographic/inventory metadata mapping definitions – which fields in the EOD files map into Alma physical/electronic inventory definitions

Metadata import files – Alma enables defining Import Profiles that define all aspects of metadata import process, including:

- Import process definitions (FTP location, scheduling, normalization, validation, matching)
- Bibliographic/inventory metadata mapping definitions – which fields in the EOD files map into Alma physical/electronic inventory definitions

Metadata imports may also be done singularly, using tools such as OCLC Connexion.

Invoice EDI import – Alma defines parameters for importing invoices from EDI files, based on Integration Profiles that define when and from where the EDI files will be imported.

Batch Updates

Batch update actions may be performed using Alma's Process Automation tools. These tools enable:

- Defining a set of records to work on. These records may be of various types:
  - Title records
  - Physical title records
  - Electronic title records
  - Physical item records
  - Digital file metadata records
- Defining a chain of defined tasks to be run on the set. These chains include:
  - Normalization of the set's records
  - Global change of record information

Are there any governance thresholds or restrictions for the import and export of data?

As with every multi-tenant solution, Alma has Governance Thresholds in place. Governance thresholds ensure that no single institution negatively impacts other Alma institutions, prevent performance degradation and can help reduce the risks of malicious attacks. With respect to importing and exporting of data to/from Alma, we have the following governance threshold. There are no additional costs that are involved in the exporting an importing of data from/to Alma.

The table below applies to fairly large jobs. Small jobs tend to fluctuate in times due to initialization and finalization overheads.

Usually jobs start close to submission time, but there might be occasions when jobs have a 'pending' status for up to 30 minutes due to the scheduling algorithm.

	Peak (working day) - records per hour	Off-Peak (night) – records per hour
Export	0.3M	1.2M
Manipulate	0.15M	0.6M
Import (**match by system ID**)	0.15M	0.6M

Are there any planning guidelines for batch jobs?

As a multi-tenant cloud based solution, Alma has a sophisticated batch job management architecture, which takes into account various factors such as the types of jobs running, the time of the day and the general load of the system, all in order to provide you with the services you require.

When using Alma's batch job services for managing your repository, it is useful to know how long processes can be expected to run. The guidelines provided below are based on Alma's actual production use of these services. Ex Libris

expect that, on average, the times given will not be less than listed, and often even better. This depends on the system load at any given time.

Basic Service	Peak (working day) - records per hour	Off-Peak (night) – records per hour
Metadata Import Note: The different profiles of metadata import may influence the time the service takes. For example, an EOD import profile will on average take longer than a bibliographic only update.	30k	120k
Metadata Export	150k	600k
Global Changes – MARC Bibliographic Normalization	50k	200k

What data can be instantaneously loaded?

Alma supports the ability to load data to be reflected immediately to the end user. Such data may include metadata records, attachments, single user record or EOD as well as data updates using Alma APIs to update users records, Bibs, acquisitions data and more as supported by our APIs. The performance of the load depends on several factors including the local institution network and size of the loaded data and as such the time it may take to load such data may vary form on customer to another. In general with over 230 institutions live using Alma, we are seeing very good performance of the system for importing and updating data via APIs. When there is a need to load large amounts of data, it is, typically handled as a batch job that is not reflected immediately in the system.

What is the average processing time for batch loads?

The processing times of batch files can widely range between very fast loading and processing of a few seconds to hours depending on the size of the file.

Alma makes use of an advanced jobs processing mechanism to ensure high performance and load balancing between jobs. Prioritizing processing of data in Alma is controlled on several levels:

1. Different jobs types are controlled with different allocation of resources (therefore they are independent of system resources). The job types are:
    a. UI short background jobs (usually indexing or saving of a record in the background)
    b. Customer batch jobs (Import, publishing etc.)
    c. Ex Libris batch jobs (such as quarterly indexing). These types of jobs use separate resources from customer jobs so the customer won't be affected by massive jobs running by Ex Libris
2. Within the above categories there is internal priority algorithm that acts as a semi- FIFO (first in first out). The system resources are divided between all jobs depending on the in queue time, job length (how much data the job needs to run on) and the job origin (was it scheduled or run manually).
3. Alma also automatically controls the jobs resources allocation. During off hours – more resources are allocated for job running.

As a reference, our system is used by very large institutions such as University of Minnesota, Manchester university, Leuven Consortia in Belgium, Orbis Cascade Alliance – consortia of 40 institutions – all are extensively using our batch load and processing of files on a daily base while reporting very good performance. We are happy to demo batch files load and processing capability to the university based on such file you can provide us or made by us, as you can understand,

the performance while is very good, may vary widely depending on the type of the job and its type.

Total views:

2838