別添３

将来の機械翻訳システムに係る調査資料

目次

１．調査の概要

現在、特許庁及び独立行政法人工業所有権情報・研修館（ＩＮＰＩＴ）では、特許情報プラットフォーム（Ｊ－ＰｌａｔＰａｔ)、中韓文献翻訳・検索システム、ワンポータルドシエ（ＯＰＤ)、高度産業財産ネットワーク（ＡＩＰＮ）といった特許情報サービスを通じて、機械翻訳システムを活用し、海外も含めた庁内外のユーザに特許情報の機械翻訳文を提供している。近年の知財活動のグローバル化に伴い、特許情報の重要性が急速に高まっているところ、特許情報の発信強化や検索環境の整備を推進する観点から、ユーザにとってより利便性の高い機械翻訳文を提供していくことが望まれている。

そこで、現状の機械翻訳の機能や特性、利用状況等を踏まえ、将来の機械翻訳システムの構成（アーキテクチャ等）のあり方として、それぞれ構成の異なる３つの構成案を提示する。

本調査では、これら３つの構成案についての特徴を精査し、より利便性の高い機械翻訳サービスを、より効率的な形で提供していくために採用すべき構成を検討する。

１．１．調査の手順

本調査の手順は以下のとおり。概要を表１．１に示す。

１．前提の整理
２．構成案洗い出し
３．比較検討を行う構成案の選定
４．構成案の整理
５．構成案の比較検討
６．調査の結果
７．要件整理資料に向けての前提の再整理

表１．１　本調査の手順

手順	概要
1	将来の機械翻訳システムの構成案作成における前提条件の整理を行う。
2	将来の機械翻訳システムの構成案の候補の洗い出しを行う。
3	将来の機械翻訳システムの構成案の候補の中から、比較検討を行う構成案の絞込みを行う。
4	絞込みを行った構成案について、より詳細なシステム構成の検討を行う。
5	構成案について、コストの観点から比較検討を行う。
6	比較検討の結果について導入時のコスト、今後の機能拡張時のコストの面から考察を行う。
7	上記の調査結果を基に要件整理資料作成に向けた前提条件の再整理を行う。

２．将来の機械翻訳システムに係る調査

２．１．前提の整理

本節では、将来の機械翻訳システムの前提となる項目を整理する。

２．１．１．システム利用者

将来の機械翻訳システムを利用するシステムは以下のとおりとする。

（１） ＯＰＤ

（２） ＡＩＰＮ

（３） Ｊ－ＰｌａｔＰａｔ（以降「ＪＰＰ」）

（４） 庁内システム

２．１．２．対象の言語、データ種別

将来の機械翻訳システムにおいて提供するサービスについて、翻訳対象となる言語と、処理対象となるデータ種別（公報、書類）の組み合わせを表２．１．２に示す。

表２．１．２　翻訳対象

#	項目	英語			中国語	韓国語
		書類		公報	公報	公報
		審査官向	その他			
1	日→英	○	○	○	-	-
2	日→中	-	-	-	-	-
3	日→韓	-	-	-	-	-
4	英→日	-	○	○	-	-
5	中→日	-	-	-	○	-
6	韓→日	-	-	-	-	○

２．１．３．性能・拡張性

将来の機械翻訳システムの性能・拡張性について、必要な条件や制限事項は以下のとおり。

（１） 性能拡張を実施する際、プロセス等の追加によるスケールアウトが可能な構成とする。

（２） ネットワーク速度は、業務ネットワーク用専用線：50Mbps（１本)、コンテンツ翻訳用専用線：50Mbps（１本)、インターネット：100Mbps（１本）とする。

（３） ネットワーク速度は、全てベストエフォートとする。

（４） 言語別の翻訳性能要件について表２．１．３に示す。

表２．１．３　言語別の翻訳性能要件

#	区分	項目	英語 (書類) 審査官向ＯＰＤ	英語 (書類) その他 ※１	英語 (公報) ※２	中国語 (公報)	韓国語 (公報)
1	対応言語	日→英	○	○	○	-	-
2		日→中	-	-	-	-	-
3		日→韓	-	-	-	-	-
4		日→ＡＳＥＡＮ	-	-	-	-	-
5		英→日	-	○	○	-	-
6		中→日	-	-	-	○	-
7		韓→日	-	-	-	-	○
8		ＡＳＥＡＮ→日	-	-	-	-	-
9	オンライン性能	同時アクセス	7	3	13	-	-
10		翻訳性能[秒]	5	60	60	-	-
11	バッチ性能	１時間当たりの処理件数	-	-	-	1,466	1,466

※１　公衆向ＯＰＤ、ＪＰＰからのアクセスを想定

※２　英語公報には、現行ＡＩＰＮにて実施されているテキスト翻訳、ＵＲＬ翻訳のトランザクションを含む

２．１．４．運用

将来の機械翻訳システムの運用について、必要な条件や制限事項は以下のとおり。

（１） 稼働監視、ジョブ管理を統合的に行える構成とする。

（２） 各サーバのバックアップ運用が可能な構成とする。

２．１．５．信頼性

将来の機械翻訳システムの信頼性について、必要な条件や制限事項は以下のとおり。

（１） １点障害によるサービスダウンが発生しない構成とする。

（２） 分散構成とする場合、１サーバがダウンした場合でも性能が 50%以上劣化しない構成とする。

２．１．６．システム処理方式

将来の機械翻訳システムの処理方式について、必要な条件や制限事項は以下のとおり。

（１） オンザフライ方式による翻訳処理が可能なこと。

（２） オンザフライ方式による翻訳処理を実施する場合、要求時の文献様式と同一の文献様式で翻訳後データのみ応答として返すこと。

（３） コンテンツ（バッチ）方式による翻訳処理が可能なこと。

（４） コンテンツ（バッチ）方式による翻訳処理を実施する場合、入力ＸＭＬを成型し、ＨＴＭＬにて原文及び翻訳文を出力すること。

（５） コンテンツ（バッチ）方式用に文献のサーバアップロード・ダウンロード機能を有すること。

（６） 翻訳前編集については、現行システムと同一の内容の処理が可能なこと。

（７） 複数の要求・入力様式に対応すること。

（８） 要求・入力様式が追加となった場合に容易に対応できること。

（９） 文字コードはＵＴＦ－８に対応していること。

（１０） 翻訳品質の向上機能を有すること。

（１１） 翻訳品質の向上タイミングは、１回／年とする。

２．１．７．セキュリティ

将来の機械翻訳システムのセキュリティについて、必要な条件や制限事項は以下のとおり。

（１） 庁内と庁外でアクセス（ＩＦ層）を分離すること。

（２） ＩＦ層は別のネットワークセグメントに設置し他サーバと分離すること。

（３） ＩＦ層にて要求を統合する場合は、フロントサーバとは別サーバとすること。

２．１．８．トータルコスト

将来の機械翻訳システムのセキュリティについて、必要な条件や制限事項は以下のとおり。

（１） サーバ構成の見積については、全サーバ物理サーバを前提として見積を実施する。

２．２．構成案洗い出し

本節では、将来の機械翻訳システム構成案の候補の洗い出しを行う。

構成案の作成に際しては、他システムからのアクセスを取り纏めるＩＦ層から、翻訳サーバに要求を出すＡＰ層までの処理部分と、ＡＰ層から機械翻訳を実施するＡＰ層までの翻訳エンジン部分とに分割してそれぞれ構成パターンを洗い出し、両者の組み合わせにより構成案の候補の洗い出しを行う。

２．２．１．処理部分

将来の機械翻訳システム構成案の候補の処理部分について、構成案の候補として挙げられるものを表２．２．１に示す。

表２．２．１　処理部分の構成案の候補

＃	構成案	内容
案１	用途分割パターン	利用者毎に特化したＡＰを作成する
案２	完全統合パターン	全て共通のＡＰを作成する
案３	翻訳方式分割パターン	オンザフライ、コンテンツ翻訳方式別にＡＰを共通化する
案４	データ種別パターン	対象データ種別（書類、公報）毎にＡＰを共通化する
案５	言語分割パターン	言語別にＡＰを共通化する
案６	言語・種別分割パターン	言語・データ種別毎にＡＰを共通化する
案７	言語・翻訳方向・種別分割パターン	言語・翻訳方向（和訳、翻訳）・データ種別毎にＡＰを共通化する

２．２．２．翻訳エンジン部分

将来の機械翻訳システム構成案の候補の翻訳エンジン部分について、構成案の候補として挙げられるものを表２．２．２に示す。

表２．２．２　翻訳エンジン部分の構成案の候補

＃	構成案	内容
案１	用途分割パターン	利用者毎に特化した翻訳エンジン層を作成する
案２	言語分割パターン	言語別に翻訳エンジン層を作成する
案３	言語・種別分割パターン	言語・データ種別毎に翻訳エンジン層を作成する
案４	言語・翻訳方向・種別分割パターン	言語・翻訳方向（和訳、翻訳）・データ種別毎に翻訳エンジン層を作成する

２．３．比較検討を行う構成案の選定

将来の機械翻訳システムの構成案を検討するにあたって、２．２で洗い出した全ての構成パターンの組合せについて比較検討を行うことは困難であるため、以下の手順で構成案の比較を実施する。

＜構成案の比較手順＞

（１）　２．２．１処理部分、２．２．２翻訳エンジン部分の構成案それぞれで比較を実施し、処理部分３案、翻訳エンジン部分１案に絞込みを行う。
なお、上記の比較の際は、「性能・拡張性」、「トータルコスト」を観点として実施する。

（２）　（１）にて決定した候補に改良を加え、比較検討を実施する構成案を決定する。

（３）　（２）にて決定した構成案３案に対し、構成案の精査、比較検討を行う。

＜構成案の選定結果＞

上記の手順に従い、絞り込み作業を行った結果を表２．３に◎で示す。

表２．３　構成案の選定結果

#			翻訳エンジン部分			
			案１	案２	案３	案４
			用途別	言語別	言語・種別	言語・翻訳方向・種別
処理部分	案１	用途別	－	－	－	－
	案２	完全統合	－	－	－	－
	案３	翻訳方式	－	－	◎	－
	案４	種別	－	－	－	－
	案５	言語別	－	－	◎	－
	案６	言語・種別	－	－	◎	－
	案７	言語・翻訳方向・種別	－	－	－	－

２．４．構成案の整理

本節では、２．３で絞り込んだ以下の３つの構成案について、その整理を行う。概要を表２．４に示す。

表２．４　構成案の整理

#	処理部分	翻訳部分
構成案１	言語分割パターン	言語・種別分割パターン
構成案２	翻訳方式分割パターン	言語・種別分割パターン
構成案３	言語・種別分割パターン	言語・種別分割パターン

２．４．１．構成案での共通の要件

構成案間で共通の要件を以下に示す。

（１）システムアーキテクチャ方針

本システムへのアクセスは、庁内からは専用線を介してＩＦ層のＷｅｂサーバに対して行われ、庁外からはインターネットを介してＩＦ層のＷｅｂサーバに対して行われる。

（２）システム構成

どの構成案についても、処理性能が求められる翻訳サーバについては、占有物理環境に配置し、それ以外については費用を考慮し共有物理環境上にサーバを配置する。

ただし、学習サーバについては、性能上共有物理環境への設置が可能だが、必要なメモリ量が2.5TBを超えており、共有物理環境にて用意可能な容量（最大1.9TB）を超えている

ため、占有物理環境に設置する。
なお、検証環境については、構成案間で差異がないため、比較検討の対象外とする。

＜共有物理環境＞
ＤＭＺセグメントにはＷｅｂサーバを設置する。また、運用管理のためのメールサーバも設置する。
ＡＰセグメントには、処理サーバ（ＡＰサーバ、バッチ管理サーバ）を言語毎に２台ずつ設置する。また、コンテンツ翻訳を行う文献のサーバアップロード・ダウンロードのためのコンテンツ管理サーバを設置する。
運用セグメントには、各サーバの状態管理やメンテナンスのための運用管理サーバ、定期的にデータバックアップを取得するためのバックアップサーバを設置する。
これらは全て冗長・分散構成とするため、各サーバ２台ずつ設置することとする。

＜占有物理環境＞
ＡＰセグメントには、翻訳言語、及び処理対象データごとに翻訳サーバを設置する。また、翻訳の品質向上のための学習サーバを設置する。
翻訳サーバは翻訳言語、及び処理対象データごとに要求性能が異なるため、それを満たすようにサーバを配置する。その際、ＯＰＤの日英書類翻訳については、審査官向けのサービスに特に処理性能が求められるため、日英書類翻訳サーバは、審査官向けとその他公衆向けとで分割して配置する。
また、学習サーバについては冗長・分散構成とし、サーバ２台を設置することとする。

＜受託者拠点＞
各サーバの状態管理やメンテナンスを遠隔で行うため、運用ＰＣを設置する。

（３）システム処理方式
機械翻訳システムの機能の一覧を表２．４．１に示す。
また、絞り込み作業時から以下の前提条件を追加する。
その他、現行システムの処理要件との関連については、「別紙２．４．１」参照。

＜追加条件＞
ＵＲＬ翻訳については、要求元にて対象のＷｅｂサイトを取得するものとする。

表２．４．１　機能一覧

#	分類	機能名	機能概要
1	共通	要求振分機能	受信した要求から処理を行うサービスを決定し、要求元システムに処理結果を返す機能（※）
2		要求管理機能	各種機能の呼び出し制御を行う機能
3		翻訳前編集機能	受信した文章を、翻訳処理ができる形式に編集する機能
4		翻訳機能	文献を文単位に分割し、翻訳サーバにて文章の翻訳を実施する機能
5		翻訳後編集機能	翻訳した文章を、受信時と同じ形式に編集し、要求元へ送信する機能
6		集計機能	翻訳時に発生した未知語や、運用保守に係る統計情報を集計する機能
7	コンテンツ（バッチ）	文献アップロード機能（画面）	特許庁運用担当者が翻訳対象の文献をアップロードする機能
8		文献ステータス表示機能(画面)	翻訳中の文献について、処理状況を表示する機能
9		文献ダウンロード機能（画面）	特許庁運用担当者が翻訳完了データやエラーリスト等をダウンロードする機能
10		納品情報出力機能	各種納品データをそれぞれの納品形式に編集して出力する機能

※ＡＰＩ　Ｇａｔｅｗａｙによって機能を代替する構成案の場合は不要

処理フロー図を図２．４．１に示す。

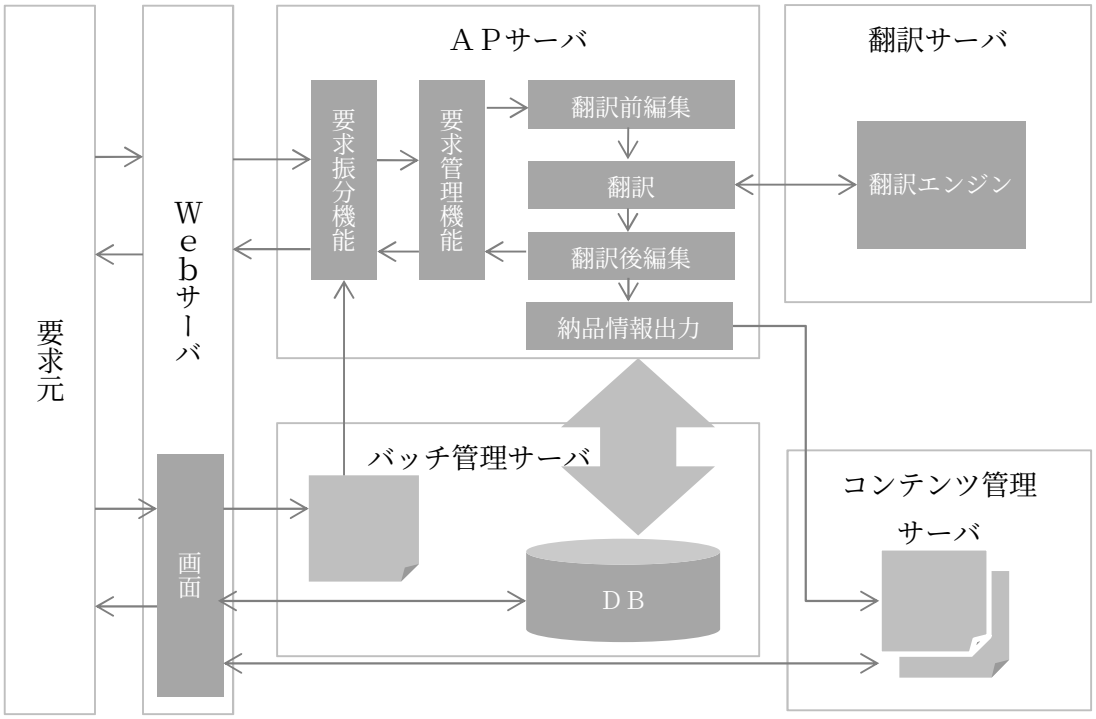


図２．４．１　処理フロー図

翻訳処理要求を受信した場合、ＷｅｂサーバからＡＰサーバへ要求が送られた後、それを受信した要求管理機能（又はＡＰＩ　Ｇａｔｅｗａｙ）が、データの種別、翻訳方向の判断を行い、それ応じた翻訳前編集機能、翻訳機能、翻訳後編集機能を随時呼び出し、完了後に処理結果を要求元へ送信する。

また、文献をサーバへアップロードしての翻訳要求の際も、同様に翻訳処理を行い、翻訳状況確認の際はその処理状況を随時更新しているバッチ管理サーバ内のステータス管理ＤＢを参照することでステータスの確認を行う。

この処理による翻訳結果は、納品情報出力機能を通じてコンテンツ管理サーバで保存し、データを取得する際はダウンロード画面より行う。

２．４．２．各構成案の特徴

各構成案の特徴を以下に示す。

（１）システムアーキテクチャ方針

各構成案について、システムアーキテクチャ方針における特徴を表２．４．２－１に示す。

表２．４．２－１　システムアーキテクチャ方針

#	処理方式
構成案１	ＡＰＩ　Ｇａｔｅｗａｙ がＡＰ層の各アプリケーションに処理を振り分ける。
構成案２	ＩＦ層に対応するＡＰ層へ処理要求を行う。
構成案３	ＡＰＩ　Ｇａｔｅｗａｙ がＡＰ層の各アプリケーションに処理を振り分ける。

（２）システム構成

各構成案について、システム構成における特徴を表２．４．２－２に示す。

表２．４．２－２　システム構成

#	Ｗｅｂサーバ	ＡＰＩ　Ｇａｔｅｗａｙ	処理サーバ	サーバ台数
構成案１	要求元毎	あり	言語毎	111
構成案２	翻訳方式毎	なし（※）	翻訳方式毎	105
構成案３	要求元毎	あり	言語・種別毎	115

※プログラムの機能追加により対応する。

（３）システム処理方式

各構成案について、システム処理方式における特徴を表２．４．２－３に示す。

表２．４．２－３　システム処理方式

#	処理方式
構成案１	ＡＰＩ　Ｇａｔｅｗａｙにより言語別に振り分けられた処理に対し、要求管理機能がデータ種別、翻訳方向を判断して、以降の処理を実施する。
構成案２	要求振分機能にて言語別に処理を振り分け、要求管理機能がデータ種別、翻訳方向を判断して、以降の処理を実施する。
構成案３	ＡＰＩ　Ｇａｔｅｗａｙにより言語・データ種別ごとに振り分けられた処理に対し、要求管理機能が翻訳方向を判断して、以降の処理を実施する。

２．４．３．各構成案の比較検討

本節では、３つの構成案に対して、それぞれ以下に示す項目について比較・検討を実施する。また、コストを算出するための翻訳エンジンの学習時や翻訳時の基礎数値は、「別紙２．４．３」参照。

＜比較項目＞

（1） 初期構築に関するコスト

システムを初期構築する際の一時経費と、その後の運用コスト、及びそれらを合計したトータルコストについて比較を行う。

（2） 性能・拡張性に関するコスト

機械翻訳システムの性能・機能の拡張を図る際に発生するコストについての比較を行う。

また、信頼性、セキュリティ、運用に関する比較は、システム構成により差異が発生しないため、ここでは比較検討の対象外とする。

（1）初期構築に関するコスト

3つの構成案について、初期構築時のコスト観点から比較した結果を以下に示す。

① サーバ台数

3つの構成案について、それぞれ初期構築時に必要となるサーバ台数を表2．4．3－1に示す。

表2．4．3－1 サーバ台数

#	物理サーバ台数	仮想サーバ台数	合計
案1	28	83	111
案2	28	77	105
案3	28	87	115

② 一時経費

3つの構成案について、想定される初期構築時の一時経費を表2．4．3－2に示す。

なお、各値は当初想定との比率（想定比）と構成案内の比率（構成比）を算出したものである。

表２．４．３－２　一時経費

#	項目	案１		案２		案３	
		想定比	構成比	想定比	構成比	想定比	構成比
1	ハードウェア （導入経費含む）	0.62	0.42	0.60	0.39	0.62	0.42
2	ソフトウェア	1.09	0.14	1.08	0.14	1.10	0.14
3	データセンタ利用費 （ＤＣ）	1.00	0.01	1.00	0.00	1.00	0.01
4	ネットワーク	1.00	0.07	1.00	0.07	1.00	0.07
5	プログラム開発	1.00	0.18	1.32	0.24	1.00	0.18
6	環境構築	1.15	0.18	1.05	0.16	1.20	0.18
7	一時経費全体の比率	0.82	-	0.84	-	0.83	-

一時経費において、構成案ごとに大きく差が見られるのは、プログラム開発、ハードウェア、ソフトウェア、環境構築の項目である。

プログラム開発は、他の構成案に比べて機能が多くなる案２のコストが高くなり、他の案が低くなっている。

一方、費用がサーバ台数に依存するハードウェア、ソフトウェア、環境構築については、サーバ台数の少ない案２のコストが低く、サーバ台数の多い案３のコストが最も高くなっている。

これらを総合した一時経費全体比率で比較すると、案２のコストが最も高く、案１のコストが最も低くなっている。

③　運用費

３つの構成案について、想定される初期構築時の運用費を表２．４．３－３に示す。

なお、各値は当初想定との比率（想定比）と構成案内の比率（構成比）を算出したものである。

表２．４．３－３　運用費

#	項目	案１		案２		案３	
		想定費	構成比	想定比	構成比	想定比	構成比
1	ハードウェア保守（共有物理含む）	0.89	0.36	0.82	0.35	0.91	0.36
2	ソフトウェア保守	1.09	0.16	1.08	0.16	1.09	0.16
3	データセンタ利用費（ＤＣ）	0.35	0.06	0.35	0.06	0.35	0.06
4	ネットワーク	1.00	0.17	1.00	0.18	1.00	0.17
5	稼働維持（訳質向上含む）	1.15	0.16	1.04	0.15	1.20	0.16
6	蓄積等運用費	0.43	0.09	0.43	0.10	0.43	0.09
7	運用費全体比率	0.80	-	0.77	-	0.81	-

運用費において、構成案ごとに大きく差が見られるのは、ハードウェア保守、稼働維持の項目である。

ハードウェア保守、稼動維持費用についてもサーバ台数の影響を受けるため、案３のコストが高く、案２のコストが低いという結果になっており、運用費全体比率を見ても同様の結果となっている。

④　トータルコスト

３つの構成案について、想定される初期構築時のトータルコストを表２．４．３－４に示す。

なお、各値は当初想定との比率（想定比）と構成案内の比率（構成比）を算出したものである。

表２．４．３－４　トータルコスト

#	項目	案１		案２		案３	
		想定比	構成比	想定比	構成比	想定比	構成比
1	一時経費全体比率	0.82	0.63	0.84	0.64	0.83	0.62
2	運用費全体比率	0.80	0.37	0.77	0.36	0.81	0.38
3	トータルコスト比率	0.81	-	0.81	-	0.82	-

一時経費、運用費を総合したところ、案３のコストが一番高く、他２案が横並びという結果となった。

これは前述のとおり、システムの導入に際しては、プログラム開発よりもサーバの台

数に依存する費用が多いことを受け、台数の多い案３のトータルコストが高くなる一方で、サーバ台数が少ないが一時経費のかかる案２と、サーバ台数が２番目に多くなるが一時経費が最も低い案１について、全体としてのコストが低くなっている。

（２）性能・拡張性に関するコスト

３つの構成案について、性能、及び拡張性の観点から比較した結果を以下に示す。

なお、どの構成においても処理性能は翻訳エンジンの性能に依存するため、拡張性についての比較のみ実施する。

また、性能に関する拡張性については、各処理のプロセスの増設にて対応を行う想定であることから、各構成案間で差異は発生しないため、ここでは機能拡張性について比較検討を行う。

①　機能拡張性

（ａ）機能拡張時の影響

３つの構成案について、機能拡張時に影響を受ける機能を比較した結果を表２．４．３－５に示す。

表２．４．３－５　機能拡張時の影響

#	分類	機能名	機能の比重	言語追加			種別追加			言語方向追加		
				案1	案2	案3	案1	案2	案3	案1	案2	案3
1	共通	要求振分機能	21.05	－	▲	－	－	▲	－	－	▲	－
2		要求管理機能	5.26	○	▲	○	▲	▲	○	▲	▲	▲
3		翻訳前編集機能	10.53	○	○	○	○	○	○	○	○	○
4		翻訳機能	21.05	○	○	○	○	○	○	○	○	○
5		翻訳後編集機能	10.53	○	○	○	○	○	○	○	○	○
6		集計機能	10.53	○	▲	○	▲	▲	○	▲	▲	▲
7	コンテンツ（バッチ）	文献アップロード機能(画面)	5.26	▲	▲	▲	▲	▲	▲	▲	▲	▲
8		文献ステータス表示機能(画面)	5.26	▲	▲	▲	▲	▲	▲	▲	▲	▲
9		文献ダウンロード機能(画面)	5.26	▲	▲	▲	▲	▲	▲	▲	▲	▲
10		納品情報出力機能	5.26	▲	▲	▲	▲	▲	▲	▲	▲	▲

ここでは、機能拡張のパターンとして、新規に対象言語を追加した場合（言語追加)、現在未対応のデータ種別へ対応する場合（種別追加)、現在未対応の翻訳方向へ対応する場合（言語方向追加）を挙げ、それぞれどの機能を改造（▲の部分）又は新規作成（○の部分）する必要があるかを、構成案ごとに示している。

翻訳方式により処理が統合されている案２は、言語、種別、言語方向のどの拡張パターンにおいても、既存処理に対する処理追加という形で機能拡張を実現するため、全体的に改造対象となる機能が多くなっている。

また、種別追加時には、案１は案２同様に改造が必要となってくるが、案３はオンザフライの改造が不要となっている。

案３は、言語、及び処理対象データ種別により分割された構成となっているため、種別追加の際は、既存機能を流用した新しい機能の追加という形で機能拡張を実現する一方で、案１は、言語ごとに処理が統合されているため、既存処理に対する処理追加という形で機能拡張を実現することから、改造範囲に差異が発生している。

（ｂ）機能拡張時の改造規模

３つの構成案について、機能拡張のパターンとそれによるプログラムの改造規模を比較した結果を表２．４．３－６に示す。

なおこの値は、それぞれの構成案の初期構築時の規模に対する比率となっている。

表２．４．３－６　機能拡張時の改造規模

#	案１	案２	案３
言語追加	1.00	1.58	1.00
種別追加	1.16	1.58	1.00
言語方向追加	1.16	1.58	1.16

どの機能拡張パターンにおいても、改造対象となる機能が多い案２の改造規模が大きくなっている。

また一方で、種別追加時には、案１は案２同様の改造が必要となってくるが、改造対象となる機能が案２に比べ少ないため、改造規模は抑えられている。

言語追加時、種別追加時には案３は必要な改造箇所が少ないため、改造規模が小さくなっているが、言語方向の追加時には、案１と同様の改造が必要であることから、改造規模が案１と横並びになっている。

② 機能拡張時の費用

（ａ）機能拡張時の一時費用

３つの構成案について、機能拡張を実施した場合に環境構築費用（一時費用）について、初期構築時の一時費用に対する比率を比較した結果を表２．４．３－７に示す。

ここでは、言語やデータ種別により必要となる構成が異なるため、今後対応が予想される具体的な言語、種別、言語方向等の機能拡張パターンを洗い出し、それぞれ値を算出している。

表２．４．３－７　機能拡張時の一時費用

#	拡張パターン	拡張内容	案１※	案２※	案３※
1	種別追加	英語（審査官向)、公報、和訳追加	0.59	0.73	0.55
2		英語（審査官向)、公報、翻訳追加	0.59	0.73	0.55
3		中国語、書類、和訳追加	0.49	0.62	0.45
4		中国語、書類、翻訳追加	0.49	0.62	0.45
5		韓国語、書類、和訳追加	0.42	0.55	0.38
6		韓国語、書類、翻訳追加	0.42	0.55	0.38
7	言語方向追加	英語（審査官向)、書類、和訳追加	0.45	0.58	0.45
8		中国語、公報、翻訳追加	0.40	0.53	0.41
9		韓国語、公報、翻訳追加	0.45	0.58	0.46
10	言語追加	ＡＳＥＡＮ、書類、和訳追加	0.38	0.55	0.38
11		ＡＳＥＡＮ、書類、翻訳追加	0.38	0.55	0.38
12		ＡＳＥＡＮ、公報、和訳追加	0.38	0.55	0.38
13		ＡＳＥＡＮ、公報、翻訳追加	0.38	0.55	0.38
14	その他	要求元追加	0.16	0.16	0.17

※上記の係数については、共通化等での費用削減を考慮していない最大の費用にて比率を出している。

全体として、案２の一時費用が大きくなっている。

また種別追加の際は、案１の値が大きいが、言語方向追加の場合は案３の値が大きくなっており、機能拡張時の開発規模の大きさと、サーバ増設数の多寡が反映された値となっている。

（b）機能拡張時の運用費

３つの構成案について、機能拡張を実施した場合に増加する運用費について、機能拡張を行わない場合の運用費に対する比率を比較した結果を表２．４．３－８に示す。

表２．４．３－８　機能拡張時の運用費

#	拡張パターン	拡張内容	案１※	案２※	案３※
1	種別追加	英語（審査官向）、公報、和訳追加	1.31	1.31	1.34
2		英語（審査官向）、公報、翻訳追加	1.31	1.31	1.34
3		中国語、書類、和訳追加	0.65	0.66	0.69
4		中国語、書類、翻訳追加	0.65	0.66	0.69
5		韓国語、書類、和訳追加	0.22	0.22	0.25
6		韓国語、書類、翻訳追加	0.22	0.22	0.25
7	言語方向追加	英語（審査官向）、書類、和訳追加	0.00	0.00	0.00
8		中国語、公報、翻訳追加	0.00	0.00	0.00
9		韓国語、公報、翻訳追加	0.00	0.00	0.00
10	言語追加	ＡＳＥＡＮ、書類、和訳追加	0.25	0.22	0.25
11		ＡＳＥＡＮ、書類、翻訳追加	0.25	0.22	0.25
12		ＡＳＥＡＮ、公報、和訳追加	0.25	0.22	0.25
13		ＡＳＥＡＮ、公報、翻訳追加	0.25	0.22	0.25
14	その他	要求元追加	0.02	0.00	0.02

※上記の係数については、共通化等での費用削減を考慮していない最大の費用にて比率を出している。

運用費の増加量は、サーバ増設数に依存するため、サーバ増設数の多い案３の値が全体的に大きくなっている。

（ｃ）機能拡張時の構成案ごとのトータルコスト

３つの構成案について、これまでに算出した機能拡張のパターンごとの一時費用と、運用費を合計したトータルコストを比較した結果を表２．４．３－９に示す。

表２．４．３－９　機能拡張時のトータルコスト

#	拡張パターン	拡張内容	案１※	案２※	案３※
1	種別追加	英語（審査官向)、公報、和訳追加	0.74	0.85	0.71
2		英語（審査官向)、公報、翻訳追加	0.74	0.85	0.71
3		中国語、書類、和訳追加	0.52	0.63	0.50
4		中国語、書類、翻訳追加	0.52	0.63	0.50
5		韓国語、書類、和訳追加	0.38	0.48	0.35
6		韓国語、書類、翻訳追加	0.38	0.48	0.35
7	言語方向追加	英語（審査官向)、書類、和訳追加	0.36	0.46	0.36
8		中国語、公報、翻訳追加	0.32	0.43	0.33
9		韓国語、公報、翻訳追加	0.36	0.47	0.36
10	言語追加	ＡＳＥＡＮ、書類、和訳追加	0.36	0.48	0.35
11		ＡＳＥＡＮ、書類、翻訳追加	0.36	0.48	0.35
12		ＡＳＥＡＮ、公報、和訳追加	0.36	0.48	0.35
13		ＡＳＥＡＮ、公報、翻訳追加	0.36	0.48	0.35
14	その他	要求元追加	0.13	0.13	0.14

※上記の係数については、共通化等での費用削減を考慮していない最大の費用にて比率を出している。

全体的に構成案２の費用が高くなっており、言語方向追加時には構成案１のコストが低く、言語追加、種別追加時には案３のコストが一番低くなっている。

3．調査の結果

前章までの結果から、初期構築時を機能拡張時では、機能拡張の種類によってもっともトータルコストが低くなる案が異なる結果となった。

このため、機能拡張の実施有無によって採用する構成案を選定する必要があると考える。

４．参考

参考として、以下の順番で機能拡張を実施した場合のトータルコストの推移を図４に示す。

■機能拡張順

①ＡＳＥＡＮ言語（１言語のみ）追加時

②中韓文献の書類追加時

③英語の言語方向追加時

④中韓文献の言語方向追加時

⑤要求元システム追加時

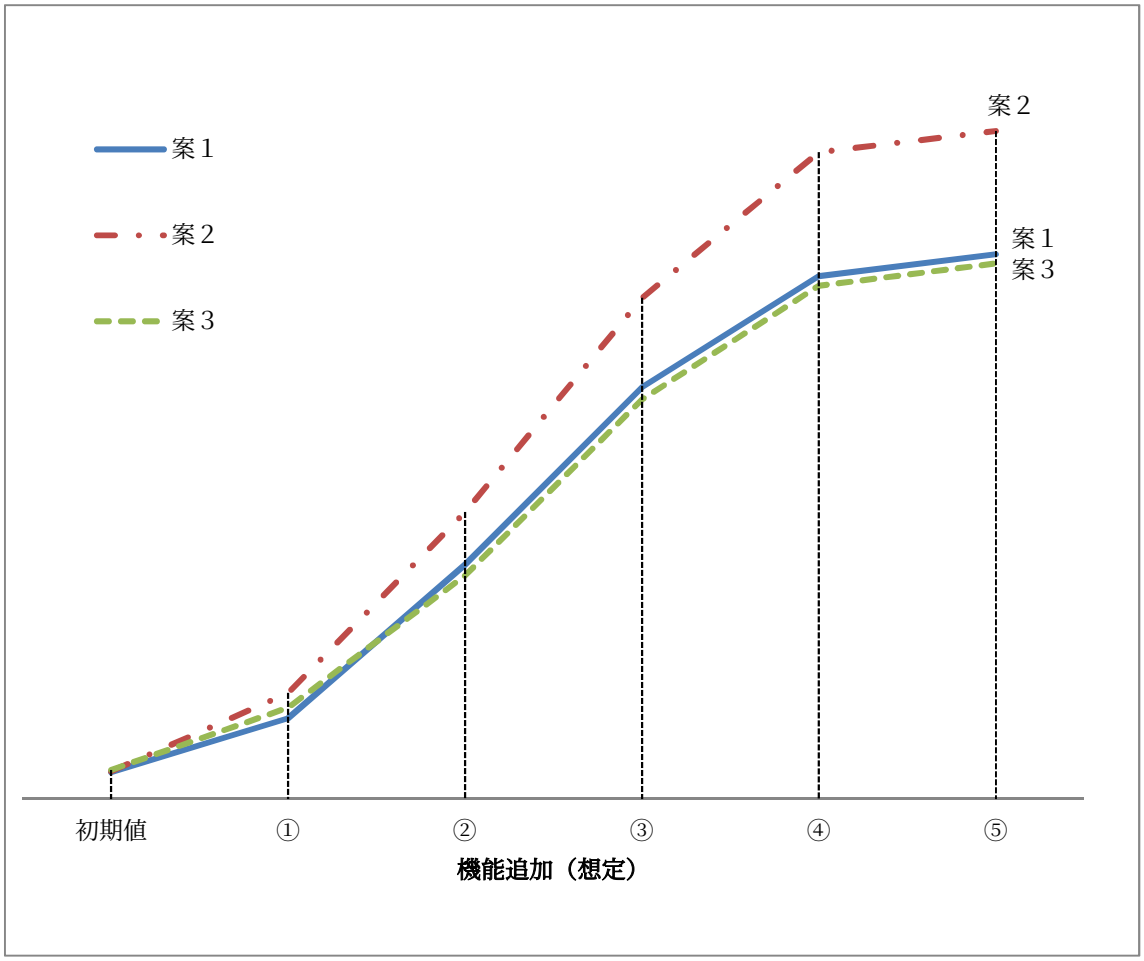


図４　トータルコストの推移

５．将来の機械翻訳システムの要件整理に向けて

本報告書では、構成の異なる３つの構成案を提示し、それぞれについての特徴を機能、コストの両面からの精査を実施した。

「４．参考」の結果から分かるとおり将来的にコストが低くなる構成は、案３の構成であるため、将来の機械翻訳システムに必用となる要件は、案３から抽出することとする。

抽出は、現行システム調査の結果と案３の構成から以下項目を抽出した。

・将来の機械翻訳システムの業務に関する条件
・将来の機械翻訳システムの機能に関する条件
・将来の機械翻訳システムの非機能に関する要件

抽出した各項目は特許庁と協議の上、現行システム調査の結果を参考に各項目を設定した。

以降に関係者間で協議し、決定した将来の機械翻訳システムの要件整理資料の前提となる項目を示す。

５．１．将来の機械翻訳システムの業務に関する条件

将来の機械翻訳システムを利用し、翻訳を実施するシステム（以下、対向システム）を表５．１－１に示す。

表５．１－１　対向システム・サービス

#	利用システム	翻訳対象データ	翻訳言語
1	ＯＰＤ	書類	日→英
2	ＪＰＰ・ＡＩＰＮ	書類	日→英
3		日本公報	日→英
4	特許庁内運用端末	中国公報	中→日
5		韓国公報	韓→日

次に、対向システムからの要求量、及び翻訳データ量算出にあたり、現行システムでの一秒あたりの同時アクセス数、及び翻訳データ量を表５．１－２、表５．１－３に示す。

表５．１－２　現行システムの同時アクセス数

#	システム名	対象サービス	同時アクセス数
1	ＯＰＤ	書類一覧翻訳	1.1 件
2		基本項目翻訳	1.2 件
3		書類実体翻訳	1.2 件
4	ＡＩＰＮ	三極向け包袋情報翻訳サービス	1.0 件
5		ＡＩＰＮ向け 包袋情報翻訳サービス	1.0 件
6		三極向け翻訳サービス	1.0 件
7		ＡＩＰＮ向け翻訳サービス	1.1 件
8	ＪＰＰ	日英公報等翻訳	1.5 件

表５．１－３　現行システムのデータ量

#	システム名	対象サービス	平均データ量		
			文字数	文数	データ量
1	ＯＰＤ	書類一覧翻訳	16 文字／文	16 文／件	8,566byte
2		基本項目翻訳	11 文字／文	38 文／件	5,885byte
3		書類実体翻訳	66 文字／文	69 文／件	12,022byte
4	ＡＩＰＮ	三極向け 包袋情報翻訳サービス	87 文字／文	64 文／件	14,429byte
5		ＡＩＰＮ向け 包袋情報翻訳サービス	70 文字／文	68 文／件	11,274byte
6		三極向け翻訳サービス	61 文字／文	940 文／件	99,972byte
7		ＡＩＰＮ向け 翻訳サービス	75 文字／文	536 文／件	93,037byte
8	ＪＰＰ	日英公報等翻訳	70 文字／文	207 文／件	35,366byte

コンテンツ翻訳を実施する中国・韓国文献の翻訳データ量は、現行中韓文献翻訳・検索システムに蓄積されている文献の発行年毎の文献数に相当するものと考えられる。

以下に、２０１０年～２０１５年にかけての中韓文献の発行件数の推移と、そこから推定される２０１６年以降の想定発行件数を表５．１－４に示す。

表５．１－４　中国・韓国文献数

公報発行年	中国	韓国
	公報発行数	公報発行数
2010 年	763,824	225,132
2011 年	946,834	252,067
2012 年	1,316,536	266,663
2013 年	1,585,129	279,434
2014 年	1,708,470	287,197
2015 年	2,099,633	252,333
2016 年	2,518,000	261,000
2017 年	2,938,000	271,000
2018 年	3,358,000	280,000
2019 年	3,778,000	290,000
2020 年	4,198,000	299,000
2021 年	4,617,000	308,000
2022 年	5,037,000	318,000
2023 年	5,457,000	327,000

予測値

また、文献毎の翻訳データ量を表５．１－５に示す。

表５．１－５　本システムのデータ量想定

#	対象サービス		平均データ量		
			文数	文字数	データ量
1	中国文献	中日翻訳	100 文字／文	110 文／件	90,000byte
2	韓国文献	韓日翻訳	110 文字／文	60 文／件	80,000byte

５．２．将来の機械翻訳システムの機能に関する条件

将来の機械翻訳システムに必要となる機能の一覧を表５．２－１に示す。

表５．２－１　機能一覧

#	機能名	書類翻訳	日本公報翻訳	中国公報翻訳	韓国公報翻訳
1	要求振分機能	○		―	
2	要求管理機能	○	○	○	○
3	翻訳前編集機能	○	○	○	○
4	翻訳機能	○	○	○	○
5	翻訳後編集機能	○	○	○	○
6	集計機能	○	○	○	○
7	文献アップロード機能	―	―	○	○
8	様式チェック機能	―	―	○	○
9	文献ステータス表示機能	―	―	○	○
10	文献ダウンロード機能	―	―	○	○
11	納品情報出力機能	―	―	○	○
12	翻訳品質向上機能	○	○	○	○

対向システムとの通信の際に送信される情報を表５．２－２に示す。

表５．２－２　インタフェース概要

#	システム名	設定内容		
		対象データ・種別（※１）	翻訳前言語	翻訳後言語
1	ＯＰＤ	書類	日本語	英語
2	ＪＰＰ	書類	日本語	英語
3		日本公報	日本語	英語
4	特許庁内運用端末	中国公報	中国語	日本語
5		韓国公報	韓国語	日本語

※１　翻訳対象の書類名、文献種別が判断可能な値が設定される。

５．３．将来の機械翻訳システムの非機能に関する要件

５．３．１．翻訳精度

将来の機械翻訳システムにおいて採用する機械翻訳エンジンについて、求められる翻訳精度を表５．３．１に示す。

表５．３．１　翻訳エンジンに求められるＢＬＥＵ等

#	翻訳前言語	翻訳後言語	データ種別			
			書類		公報	
			ＢＬＥＵ	ＲＩＢＥＳ	ＢＬＥＵ	ＲＩＢＥＳ
1	日本語	英語	25	70	25	70
2	中国語	日本語	-	-	25	80
3	韓国語	日本語	-	-	55	90

５．３．２．翻訳性能

対向システムからの要求を受けてから、翻訳を実施し、翻訳文を送信するまでに必要な時間について、求められる翻訳性能を表５．３．２－１に示す。

表５．３．２－１　性能要件

#	システム	翻訳対象データ（※）	言語	性能
1	ＯＰＤ	書類一覧	日→英	２秒/要求
2		書類実体	日→英	８秒/要求
3	ＪＰＰ	書類一覧	日→英	２秒/要求
4		書類実体	日→英	８秒/要求
5		日本公報	日→英	30 秒/要求
6	特許庁内運用端末	中国公報	中→日	800 件/時
7		韓国公報	韓→日	50 件/時

※　各データの文数、文字数については、「５．１将来の機械翻訳システムの業務に関する要件」参照

また、対向システムとのネットワーク回線について、求められる性能を表５．３．２－２に示す。

表５．３．２－２　ネットワーク回線一覧

#	回線	システム	速度
1	業務ネットワーク用専用線	ＯＰＤ	50Mbps
2	インターネット回線	ＪＰＰ 特許庁内運用端末	100Mbps

※速度はベストエフォートでの値とする。

別紙２．４．１

本資料は、「比較検討を行う構成案の選定」後に「構成案の整理」を行うにあたり、各構成案に必要となる機能等を整理したものである。

各構成案について詳細なシステム構成の検討を行うため、「前提の整理」の際に整理した最低限のシステム処理方式を元に、必要な機能要件を比較検討が可能なレベルまで具体化することを目的とする。

１．現行システム分析

将来の機械翻訳システムの基となる現行システムは、「現行システムの機械翻訳に係る調査」にて調査を実施した以下４システムである。

本章では、現行システムの機能及び処理方式から、将来の機械翻訳システムの処理要件を抽出する。

<現行システム>

・Ｊ－ＰｌａｔＰａｔ(以降「ＪＰＰ」)
・ＯＰＤ
・ＡＩＰＮ
・中韓文献翻訳・検索システム(以降「中韓Ｓ」)

別紙２．４．１

１．１．現行システムにて提供している機能

現行システムにて提供している機能とその概要を表１．１に示す。

表１．１　現行システムの機能

#	現行システム名	機能	機能概要
1	ＪＰＰ	公報翻訳	公報データを翻訳する
2	ＯＰＤ	包袋情報翻訳	包袋情報を翻訳する
3	ＡＩＰＮ	公報翻訳	公報データを翻訳する
4		包袋情報翻訳	包袋情報を翻訳する
5		テキスト翻訳	任意のテキストデータを翻訳する
6		ＵＲＬ翻訳	任意のＷｅｂページを翻訳する
7	中韓Ｓ	公報翻訳	公報データを翻訳する
8		テキスト翻訳	任意のテキストデータを翻訳する
9		文献提供	翻訳文データの納品を行う

１．２．現行システムの処理方式

現行システムの調査結果から、類似している処理方式ごとに機能を整理した結果を表１．２に示す。

表１．２　現行システムの調査結果

#	機能	処理概要	現行システム
1	テキスト翻訳	要求を受け任意のテキストを翻訳エンジンにて翻訳を実施し、要求元に応答を返す	ＡＩＰＮ、中韓Ｓ
2	包袋情報翻訳	要求を受け包袋の様式に応じた翻訳前編集、翻訳エンジンによる翻訳、翻訳後編集の順で処理を行い、要求元に応答を返す	ＯＰＤ、ＡＩＰＮ
3	ＵＲＬ翻訳	要求を受け対象のＵＲＬからＷｅｂページを取得後、翻訳前編集、翻訳エンジンによる翻訳、翻訳後編集の順で処理を行い、要求元に応答を返す	ＡＩＰＮ
4	公報翻訳	クライアント又はサーバから要求を受け公報の様式に応じた翻訳前編集、翻訳エンジンによる翻訳、翻訳後編集の順で処理を行い、要求元に応答を返す	ＪＰＰ、ＡＩＰＮ、中韓Ｓ
5	文献提供	外部提供用の翻訳文データのみの抽出と共通庁内システムへのデータ提供用の翻訳文に対応する原文、図面イメージの抽出を行う	中韓Ｓ

別紙２．４．１

１．３．現行システムの処理方式から処理要件の抽出

現行システムにて実施している処理の概要と機能の対応を表１．３－１に示す。

表１．３－１　現行システムの処理概要

#	機能／処理	テキスト翻訳	包袋情報翻訳	ＵＲＬ翻訳	公報翻訳	文献提供	処理概要
1	要求応答	○	○	○	○	-	クライアント又は、サーバからの要求の受付及び応答を行う
2	Ｗｅｂページ取得	-	-	○	-	-	対象のＷｅｂページを取得する
3	翻訳前編集	-	○	○	○	-	様式に応じた翻訳前編集を行う
4	翻訳	○	○	○	○	-	言語に応じた翻訳を行う
5	翻訳後編集	-	○	○	○	-	様式に応じた翻訳後編集を行う
6	外部提供	-	-	-	-	○	翻訳文のみの抽出を行う
7	データ提供	-	-	-	-	○	翻訳文に対応する原文、図面イメージの抽出を行う

整理結果から、機能単位で共通性の高い処理を将来の機械翻訳システムの「必須要件」とする。

また、共通性の低い処理については、実現性及び必要性を検討した上で要件化する「任意要件」とする。

必須要件と任意要件の一覧を表１．３－２に示す。

表１．３－２　必須要件と任意要件

#	区分	処理	処理概要
1	必須	要求応答	クライアント又は、サーバからの要求の受付及び応答を行う
2		翻訳前編集	様式に応じた翻訳前編集を行う
3		翻訳	言語に応じた翻訳を行う
4		翻訳後編集	様式に応じた翻訳後編集を行う
5	任意	Ｗｅｂページ取得	対象のＷｅｂページを取得する
6		外部提供	翻訳文のみの抽出を行う
7		データ提供	翻訳文に対応する原文、図面イメージの抽出を行う

２．将来の機械翻訳システムにて採用する処理方式

本章では、現行システムから抽出した処理要件を基に将来の機械翻訳システムにて採用する処理方式の検討を行う。

２．１．将来の機械翻訳システムの位置付け

将来の機械翻訳システムは審査官向ＯＰＤ(現ＯＰＤ)、公衆向ＯＰＤ(現ＡＩＰＮ)、庁内システム(※)、ＪＰＰからの翻訳要求を受け取った後、それぞれの要求に合わせた翻訳処理を行い、応答を返すシステムとなる。要求元のシステムとの関係を図２．１に示す。

※現行の中韓Ｓについては、照会機能・検索機能が共通特実検索システム(文献照会部分、検索部分)に、翻訳機能が将来の機械翻訳システムに統合される予定となっている。

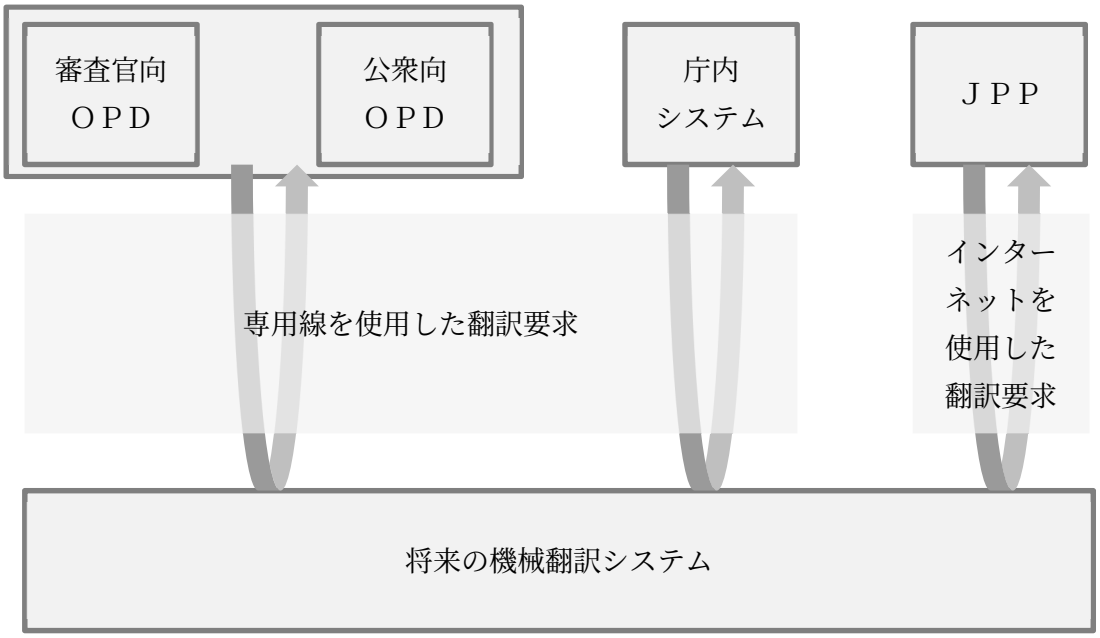


図２．１　将来の機械翻訳システムの位置付け

２．２．処理方式の検討方針

処理方式の検討については、以下方針にて採用可否を判断する。

＜検討方針＞

・必須要件については、そのまま削除可否の検討は行わない
・１機能、１サービスとなるように細分化する
・任意要件については、要求元で処理することで開発規模以外のメリットがある場合に、メリット・デメリットを整理する

２．３．任意要件の精査

現行システムの調査資料から抽出した処理要件において、任意要件となった項目の精査を行う。

以下に任意要件を要求元システムにて実施した場合のメリット・デメリットを示す。

（１）Ｗｅｂページ取得

ＵＲＬ翻訳要求にて指定されたＷｅｂページの取得を行う処理を要求元で実施した場合のメリット・デメリットを以下に示す。

＜メリット＞

・将来の機械翻訳システム側のセキュリティ攻撃の脅威を抑止できる

→将来の機械翻訳システムにて処理を実施した場合、本来のアクセスパスが特許庁、ＪＰＰのみのところ、インターネット全てにパスを開く必要があるため、セキュリティ攻撃を受けるリスクが高まる。

＜デメリット＞

・特に無し

＜考察＞

ＵＲＬ翻訳要求にて指定されたＷｅｂページの取得を行う処理を要求元で実施した場合のデメリットは無いと考える。

さらに、将来の機械翻訳システム側で実施した場合、セキュリティリスクが高まるため、セキュリティ要件のレベルが上がり、更なるコスト増につながると考える。

このため、本要件については、将来の機械翻訳システムの処理要件とせず、要求元が対応することを前提とする。

（２） 外部提供

翻訳文のみの抽出を行う処理を、要求元で行う場合のメリット・デメリットを以下に示す。

＜メリット＞

・ユーザへの情報提供の速度が向上する

→将来の機械翻訳システムの受託者からの抽出結果の提出を待たずに、処理が出来ている部分までの外部提供用の媒体（アーカイブ）作成が可能となる。

＜デメリット＞

・要求元側での作り込みが必要

→現状、要求元がシステム化されていないため、抽出、外部提供のためのシステム化、又は運用の追加が必要となる。

・要求元側の運用が増加する

→現状､中韓Ｓ側にて外部提供用の媒体を作成している運用が将来の機械翻訳システムの稼働後は、要求元システム側の運用となる。

＜考察＞

翻訳のみの抽出を行う処理を要求元で行う場合、メリットに対してデメリットが多く、要求元の運用負荷が上がる可能性がある。

このため、本要件については、将来の機械翻訳システム側で実施する要件とする。

ただし、デメリットを許容するほどメリットの効果が大きいと判断される場合は、要求元が対応することを前提として要件を再整理するものとする。

(３) データ提供

翻訳文に対応する原文、図面イメージの抽出を行う処理を要求元で行う場合のメリット・デメリットについては、「(２)外部提供」と同一であると考える。

このため、「(２)外部提供」の対応と同様に将来の機械翻訳システム側で実施する要件とする。

また､デメリットを許容するほどメリットの効果が大きいと判断される場合の対応についても「(２)外部提供」と同様に対応する。

２．４．処理方式の実現性検討

ここまでに整理した処理方式について､将来の将来の機械翻訳システムにて使用した場合の課題とその対策を以下に示す。

＜現状想定する処理方式の課題＞

現行の機械翻訳システムでは、システム毎に翻訳機能が存在しているため、要求元機能と翻訳機能が１対１の関係となっているが、将来の機械翻訳システムでは、複数システムから複数の機能に対して翻訳要求の受付・応答を行う必要がある。

＜対策＞

要求応答の処理については、これまで通り要求を受け付けて応答を返す処理に加えて、要求元システムと要求内容から対応する処理を判断して､要求を振り分けるための処理要件を追加することで対応する。

また、これを受けての将来の機械翻訳システムの処理方式は以下のとおりとする。

別紙２．４．１

＜課題対策後の処理方式＞

前提：ＵＲＬ翻訳要求における翻訳対象のＷｅｂページについては、要求元システム側にて取得されるものとする。

処理概要を表２．４に示す。

表２．４　処理概要

#	処理	処理概要
1	要求振分	要求元に対応する処理へ振分を行う
2	要求応答	クライアント又は、サーバからの要求の受付及び応答を行う
3	翻訳前編集	様式に応じた翻訳前編集を行う
4	翻訳	言語に応じた翻訳を行う
5	翻訳後編集	様式に応じた翻訳後編集を行う
6	外部提供	翻訳文のみの抽出を行う
7	データ提供	翻訳文に対応する原文、図面イメージの抽出を行う

別紙２．４．３　基礎数値取得結果

第一部　多重度１

第二部　多重度２～

第一部　多重度１

取得対象

学習時

学習所要時間, メモリ使用量, CPU使用率
ディスク使用量　(ディスクI/O)

翻訳時

翻訳サーバ起動時間(メモリ展開時間)
翻訳レスポンス時間, TPM, メモリ使用量, CPU使用率
(ディスクI/O)

実行環境

クラウド：Amazon AWS r3.8xlarge
(Intel Xeon E5-2670 v2プロセッサー 32vCPU, 244GB RAM, 600GB SSD)
OS：CentOS release 6.5　64bit

取得条件

学習時 (すべて多重度1†)

- 日⇔英：1000万コーパス, 100万コーパス
- 日⇔韓：100万コーパス
- 日⇔中：100万コーパス

(†)NICT学習ツールのdefault設定を多重度1と定義
(マルチスレッドによる並列処理を含む; max4スレッド)

翻訳時(すべて多重度1)

- 日⇔英：1000万コーパス, 100万コーパス
  日→英：30,60文字　英→日：16,32単語(‡)
- 日⇔韓：100万コーパス
  日→韓：30,60文字　韓→日：30,60文字(‡)
- 日⇔中：100万コーパス
  日→中：30,60文字　中→日：20,40文字(‡)

(‡) X→日方向の入力文長は 30文字,60文字の日本語文に
対応する当該言語の単語数or文字数に基づいて決定

取得結果／学習時

	日→英1M	日→英10M	英→日1M	英→日10M
学習所要時間	22時間34分	208時間33分	28時間27分	244時間44分
メモリ使用量(最大)	53.6GB	222.52GB	72.92GB	225.61GB
CPU使用率(平均)	11.55%	9.53%	15.30%	11.09%
ディスク使用量(終了時)	8.74GB	72.74GB	9.01GB	72.43GB
ディスク使用量(最大)	19.66GB	188.98GB	21.56GB	207.95GB

	日→中1M	日→韓1M	中→日1M	韓→日1M
学習所要時間	30時間03分	26時間23分	45時間40分	25時間44分
メモリ使用量(最大)	68.62GB	78.21GB	70.02GB	69.24GB
CPU使用率(平均)	12.87%	10.48%	23.74%	10.47%
ディスク使用量(終了時)	12.19GB	28.06GB	11.00GB	27.07GB
ディスク使用量(最大)	35.57GB	62.83GB	34.42GB	58.04GB

※ディスクI/O(read/write)も計測したが(→p.7)、遅延書き込みの影響が大きいために瞬時値が意味をなさないこと、I/Oがボトルネックとなる事象は見られなかったことから割愛した

考察／学習時

学習所要時間

- コーパスサイズに対しほぼリニアに増加
- 日英< 日韓　　韓日 < 英日 < 日中 < 中日

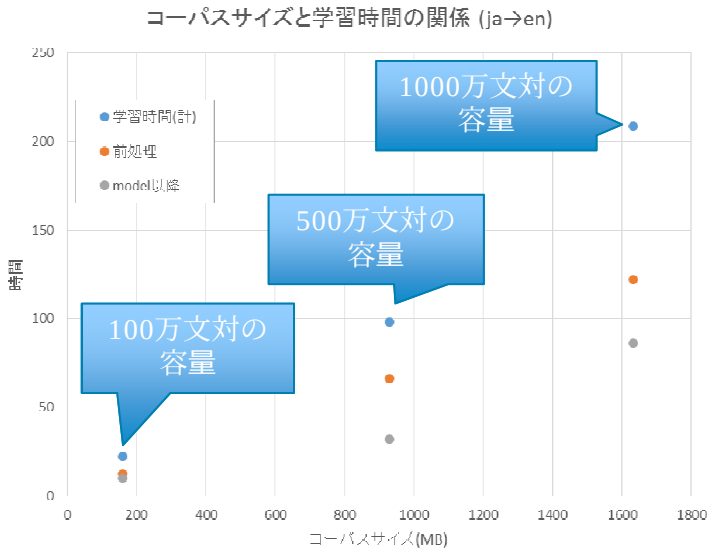
※前処理部の計算量が 日≒韓<英<中 であることによる

ディスク使用量

- コーパスサイズに対しほぼリニアに増加(終了時,ピーク時とも)

メモリ使用量(max), CPU使用率

- 言語対による違いは小



取得結果／翻訳時

	日→英1M		日→英10M		英→日1M		英→日10M	
入力	30文字	60文字	30文字	60文字	16単語	32単語	16単語	32単語
サーバ起動時間	17.1s (4.95s)		145s (5.91s)		19.9s (7.02s)		171s (7.95s)	
翻訳ﾚｽﾎﾟﾝｽ時間	88.83ms	176.06ms	104.07ms	205.32ms	112.20ms	254.45ms	126.36ms	288.25ms
TPM	675.4	340.8	576.5	292.2	534.8	235.8	474.8	208.2
メモリ使用量	4.86GB	5.57GB	24.59GB	25.23GB	6.30GB	6.47GB	29.06GB	29.70GB
CPU使用率(平均)	5.2%	5.3%	5.0%	5.2%	6.6%	8.0%	6.4%	8.1%

	日→中1M		日→韓1M		中→日1M		韓→日1M	
入力	30文字	60文字	30文字	60文字	20文字	40文字	30文字	60文字
サーバ起動時間	25.6s (5.00s)		10.2s (3.99s)		24.3s (7.65s)		6.54s (2.28s)	
翻訳ﾚｽﾎﾟﾝｽ時間	87.70ms	174.08ms	47.15ms	80.25ms	121.44ms	242.82ms	37.48ms	48.44ms
TPM	684.2	344.7	1272.5	747.7	494.1	247.1	1600.9	1238.6
メモリ使用量	6.28GB	6.90GB	2.89GB	2.91GB	7.60GB	7.81GB	1.95GB	1.96GB
CPU使用率(平均)	4.8%	5.1%	6.1%	6.1%	8.4%	9.1%	6.0%	6.1%

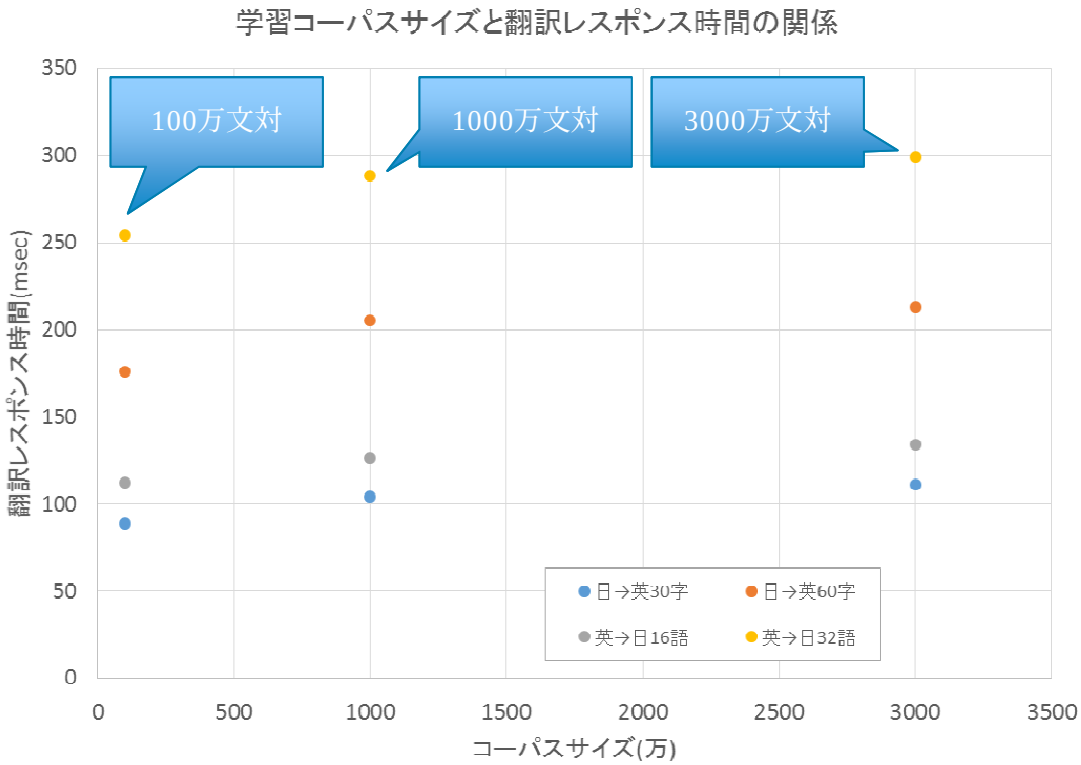
※サーバ起動時間はモデルがすべてRAMに展開されるまでの時間(カッコ内は起動の時点ですでにキャッシュにロードされている場合)を示す。

考察／翻訳時

翻訳レスポンス時間

- 入力文長に対してほぼリニアに増加
- コーパスサイズに対する変化は1000万以上では少ない

↑商用モデル(推定3000～4000万コーパス)にて実測



考察／翻訳時

メモリ使用量

- メモリ使用量 ＝ モデルサイズ ＋ α

  α　3～4GB(日⇔英中)，1～2GB(日⇔韓)

  ※コーパスサイズ増加に伴うメモリ使用量の増加はモデルサイズの増加分に一致

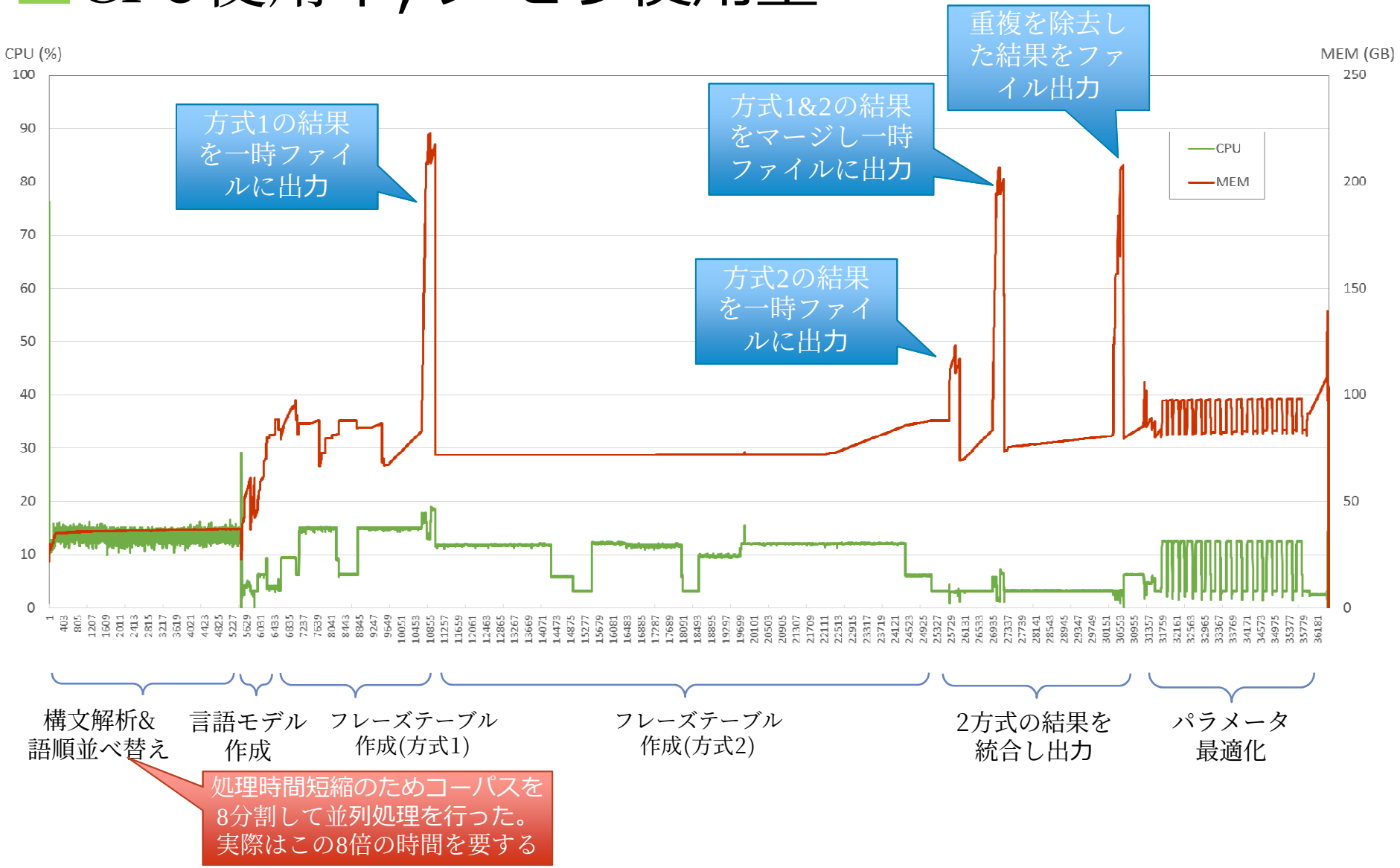
- 韓は英中に比べメモリ使用量が少ない

	日→英1M	日→英10M	英→日1M	英→日10M
モデルサイズ	2.3GB	21.5GB	2.8GB	25.1GB
メモリ使用量 (日:60文字, 英:32単語)	5.57GB	25.23GB	6.47GB	29.70GB

	日→中1M	日→韓1M	中→日1M	韓→日1M
モデルサイズ	3.6GB	0.9GB	3.3GB	1.0GB
メモリ使用量 (中:40文字, 韓:60文字)	6.90GB	2.91GB	7.81GB	1.96GB

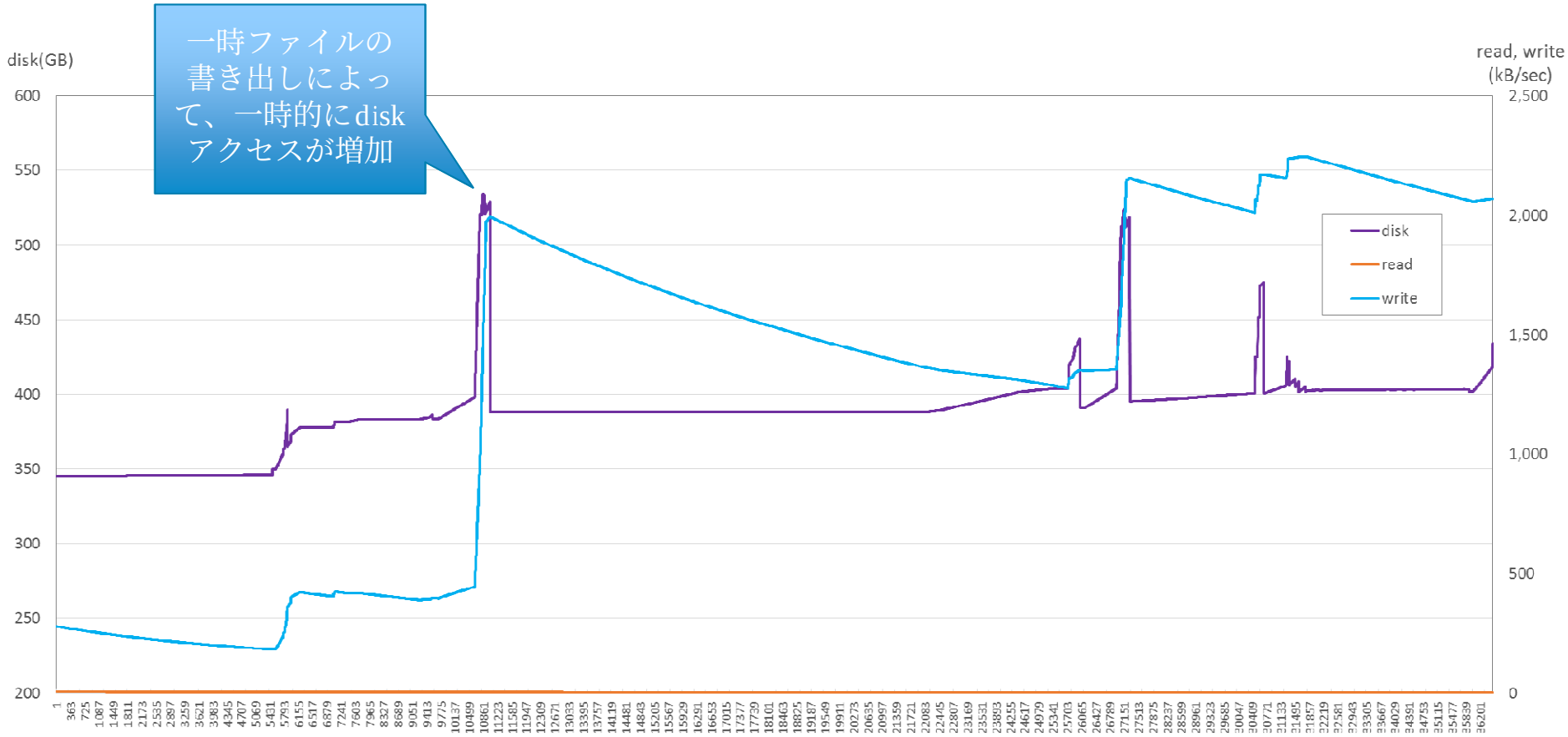
基礎数値の推移の例（学習時;ja→en 1000万）

CPU使用率, メモリ使用量



基礎数値の推移の例（学習時;ja→en 1000万）

ディスク使用量, ディスクI/O(read&write)



第二部　多重度２～

取得対象

学習時

学習所要時間, メモリ使用量, CPU使用率
ディスク使用量　(ディスクI/O)

翻訳時

平均応答時間(翻訳レスポンス時間),
平均スループット(T/S), メモリ使用量,　CPU使用率

実行環境

クラウド：Amazon AWS r3.8xlarge
(Intel Xeon E5-2670 v2プロセッサー 32vCPU, 244GB RAM, 600GB SSD)
OS：CentOS release 6.5　64bit

取得条件

学習時

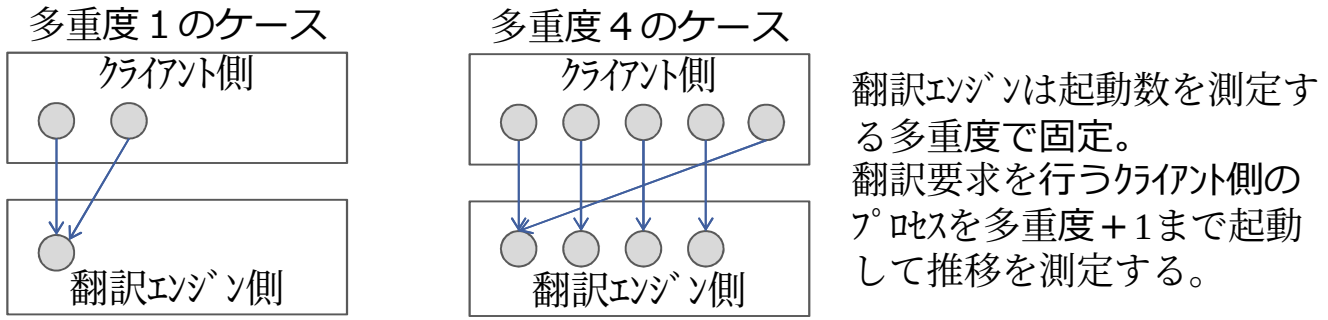
- 日⇔英：1000万コーパス
- 多重度：2, 4(†)

(†)NICT学習ツールのdefault設定を多重度1と定義
(マルチスレッドによる並列処理を含む; max4スレッド)

翻訳時

- モデル：日⇔英 1000万コーパス
- テスト文：［日→英］平均88文字, 25万文
  ［英→日］平均41単語, 25万文 (日→英テスト文の対訳)
- 多重度＝1, 2, 4, 8, 12, 16, 20, 24 ≝エンジン起動数
  クライアントプロセス数＝1～ (エンジン起動数+1)

【翻訳測定イメージ】



翻訳ｴﾝｼﾞﾝは起動数を測定する多重度で固定。
翻訳要求を行うｸﾗｲｱﾝﾄ側のﾌﾟﾛｾｽを多重度＋1まで起動して推移を測定する。

取得結果/学習時（多重度1,2,4）

日→英

	多重度１	多重度２	多重度４
学習所要時間	208時間33分	143時間36分	88時間59分
メモリ使用量(最大)	222.52GB	216.91GB	229.73GB
CPU使用率(平均)	9.53%	22.73%	34.77%
ディスク使用量(終了時)	72.74GB	80.45GB	80.87GB
ディスク使用量(最大)	188.98GB	194.44GB	196.61GB

英→日

	多重度１	多重度２	多重度４
学習所要時間	244時間44分	156時間02分	101時間17分
メモリ使用量(最大)	225.61GB	227.10GB	245.35GB
CPU使用率(平均)	11.09%	30.28%	46.39%
ディスク使用量(終了時)	72.43GB	87.45GB	80.99GB
ディスク使用量(最大)	207.95GB	209.28GB	213.49GB

考察/学習時（多重度1,2,4）

学習所要時間

- 多重度とともに短縮

  多重度2倍　→ 所要時間約 2/3

- 学習ステップによって短縮度合いが異なる（→別紙）

  ・構文解析＆語順並び替え：　多重度どおりに短縮

  ・フレーズテーブル作成(方式2)：　同上

  ・パラメータ最適化：　翻訳実行部分のみ短縮

  上記以外はシングルスレッド動作のため短縮されない（実装仕様どおり）

CPU使用率

- 多重度とともに上昇

ディスク使用量、最大メモリ使用量

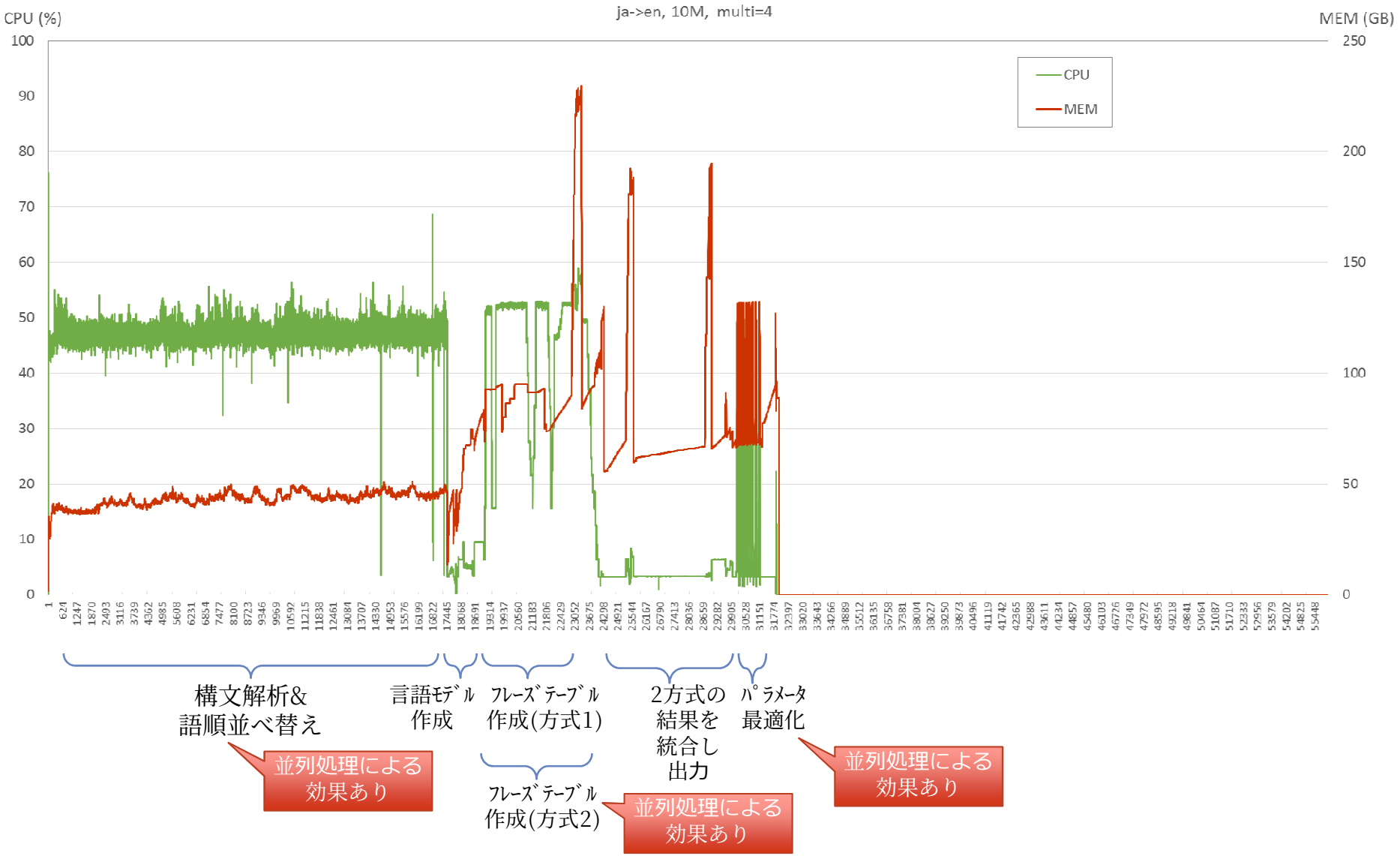
- ほぼ横ばい

基礎数値推移の例（ja→en 多重度2）

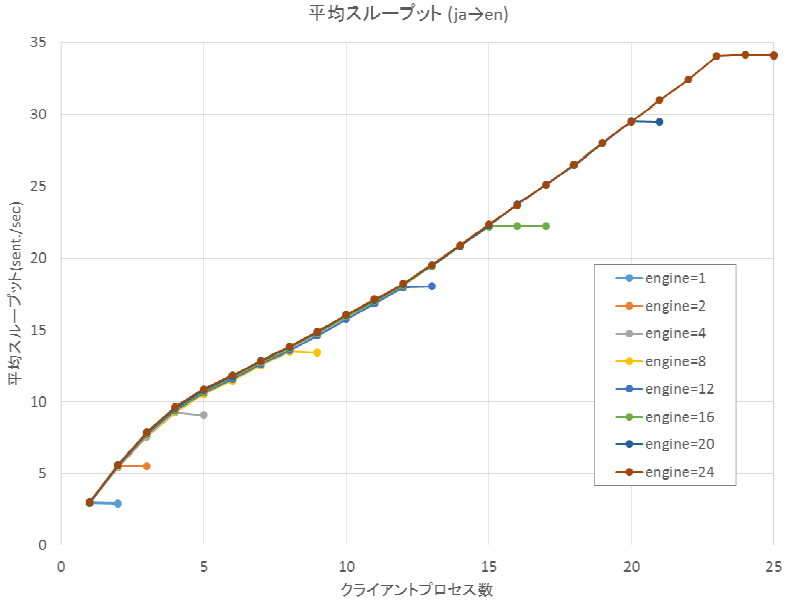
CPU使用率, メモリ使用量

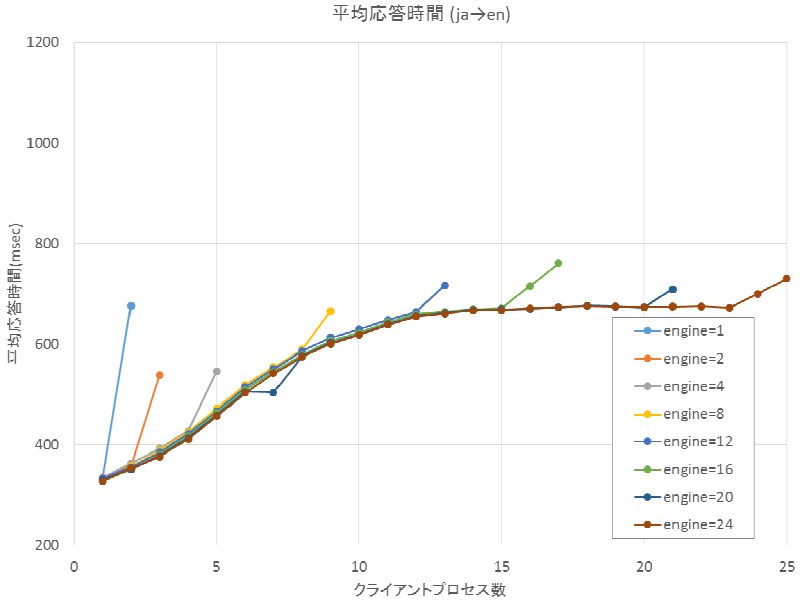
基礎数値推移の例 (ja→en 多重度4)

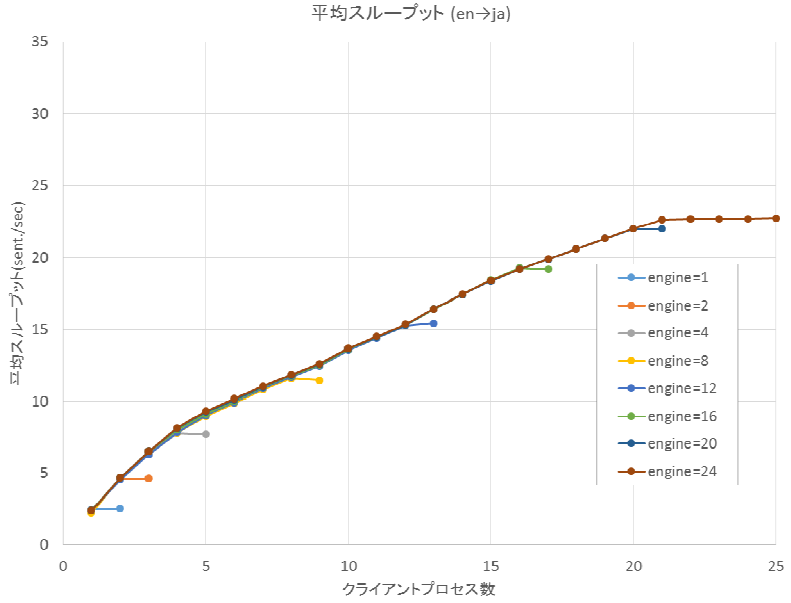
CPU使用率, メモリ使用量

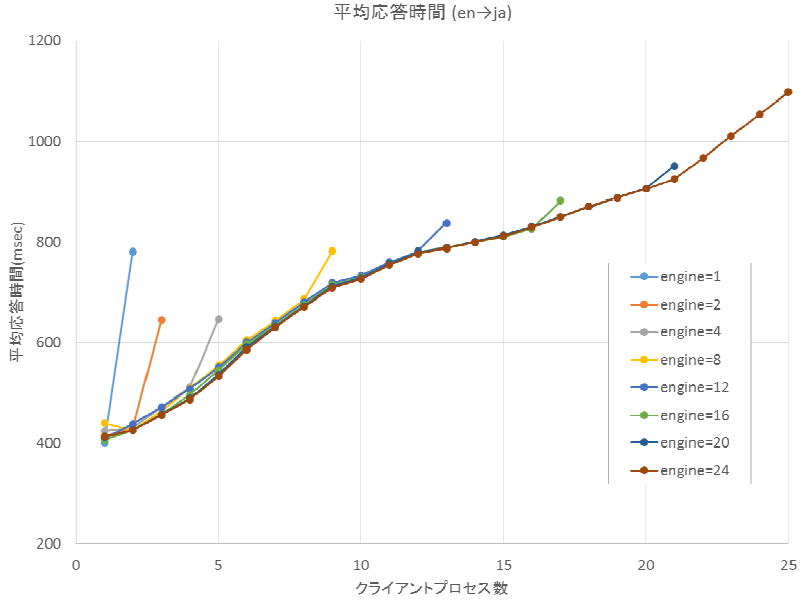


取得結果 / 翻訳時(ｽﾙｰﾌﾟｯﾄ&応答時間)









取得結果/翻訳時(ｽﾙｰﾌﾟｯﾄ&応答時間)

考察

平均スループット

・エンジン起動数N_Eを固定しクライアントプロセス数N_Cを変化

→ N_C　N_Eまでスループットが上昇

・ N_E= N_C=20のとき N_E=N_C=1の10倍

(=理想値の1/2; 日→英, 英→日とも)

・有効な多重度の上限：

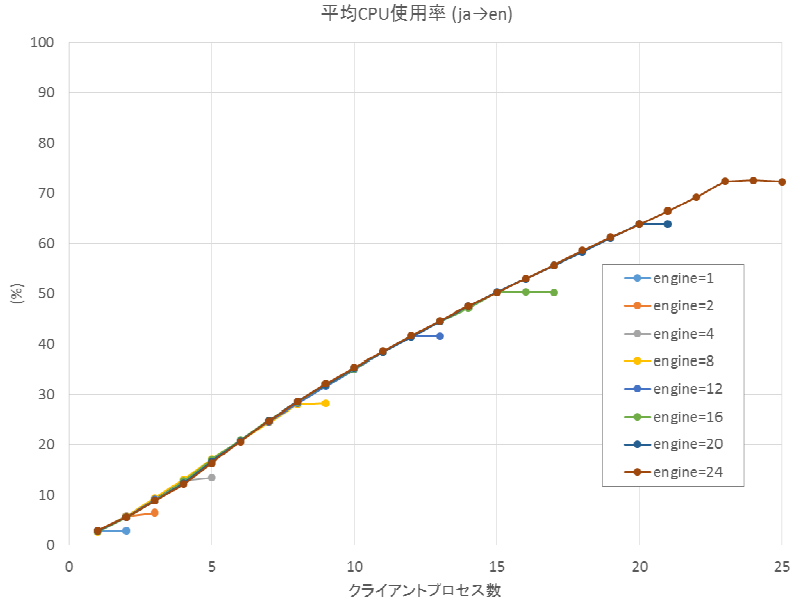
日→英　24，英→日　20 (32vCPUの場合)

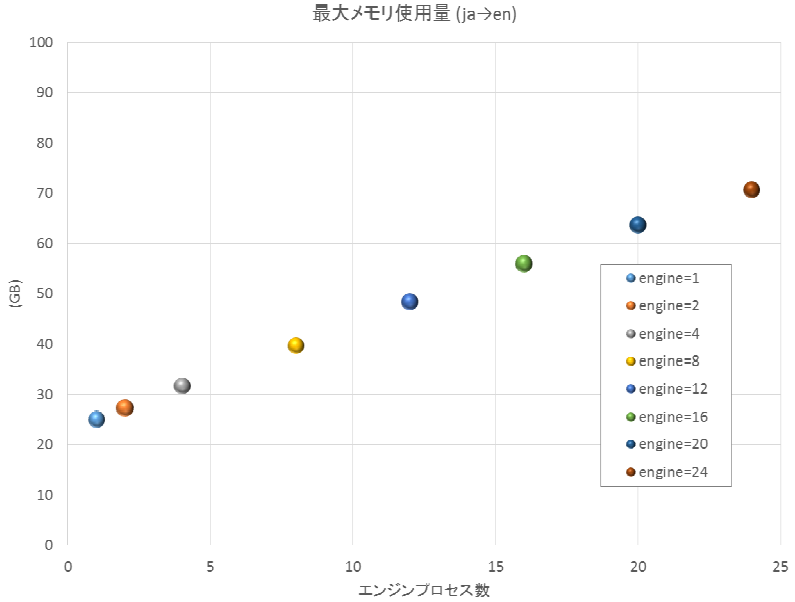
→ CPUリソース(後述)によって決まる

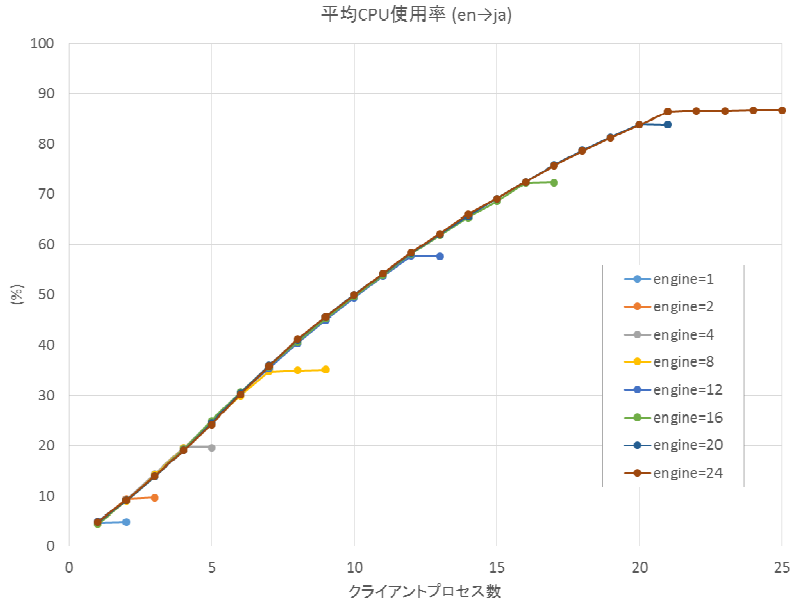
平均応答時間：

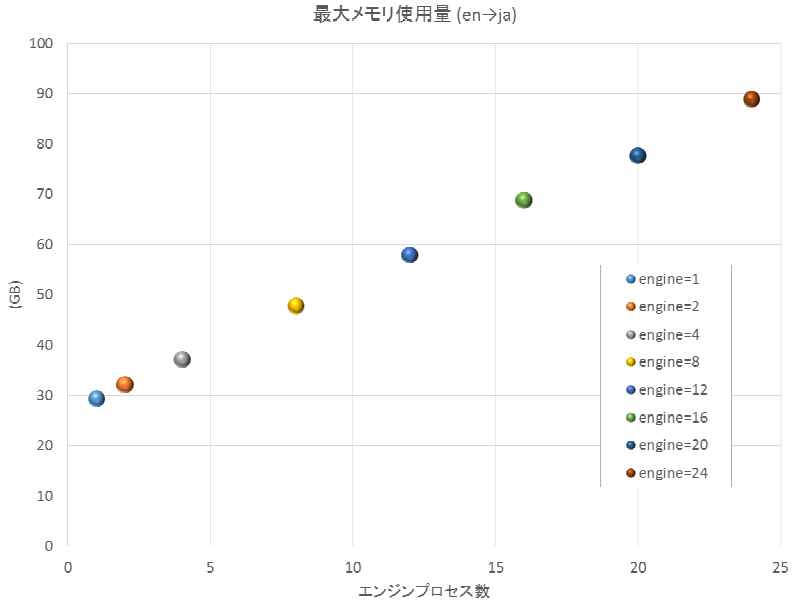
N_E= N_Cまで徐々に上昇し、N_E<N_Cから急上昇

取得結果 / 翻訳時 (CPU&メモリ)









取得結果 / 翻訳時 (CPU&メモリ)

考察

CPU使用率

・エンジン起動数N_Eを固定しクライアントプロセス数N_Cを変化

→ N_C　N_Eまでほぼリニアに増加

・80%前後で頭打ちとなる(100%近くまで使われない)

→エンジンの複雑なプロセス構成に起因すると推察

(1並列あたり30～35プロセスからなる)

メモリ使用量 ＝モデルサイズ + N_E×定数α

α　2GB(日→英)，2.5GB(英→日)

↑構文解析＆語順並び替え部分が消費

将来の機械翻訳システム構成案　案1

構成案　案1　目次

1 システムアーキテクチャ方針

1.1 アーキテクチャ概念図

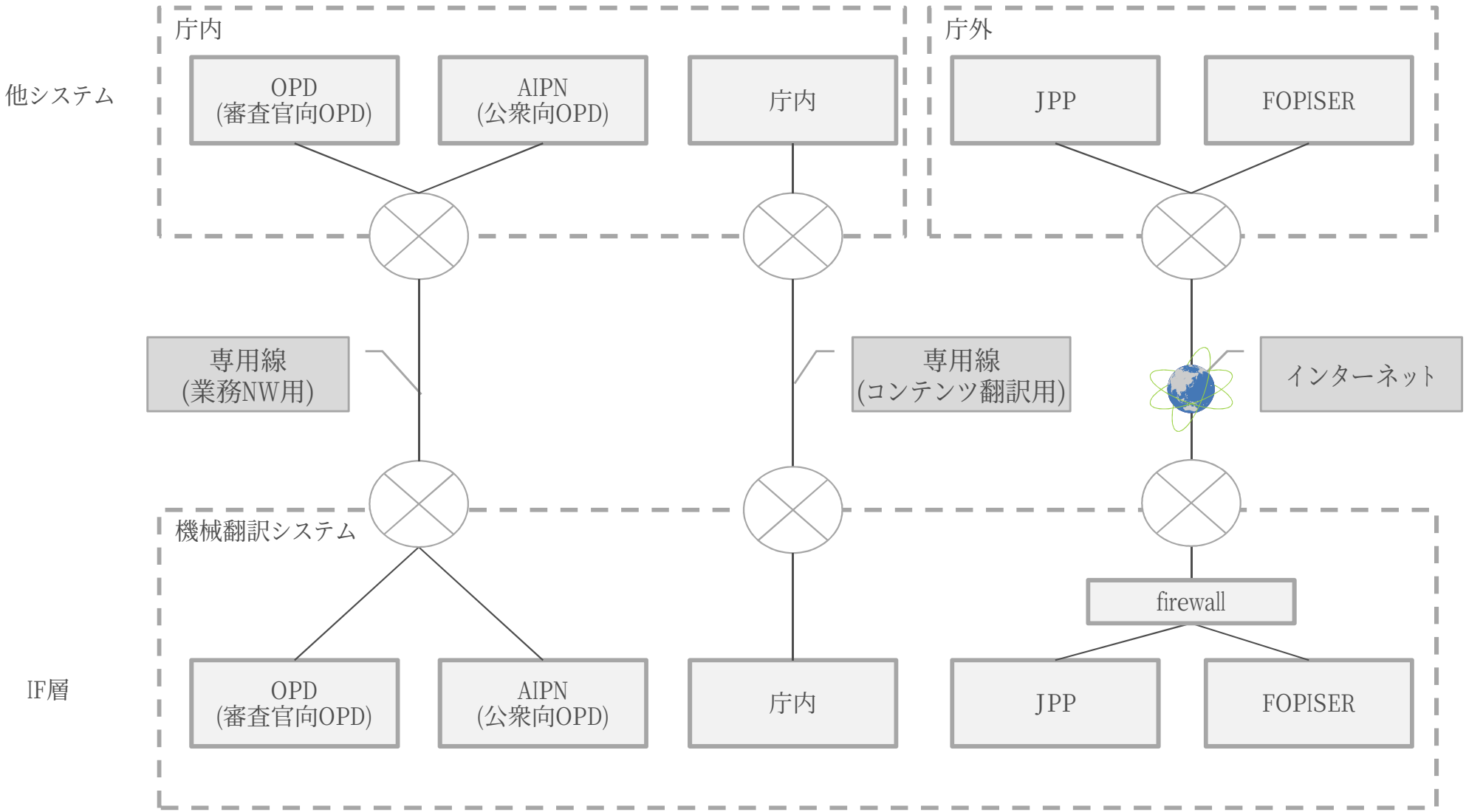
案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

案1にて採用するアーキテクチャは以下の通り。
案1では、IF層を各要求元システム分用意し、各システムから来た要求をAPI Gatewayにて一括管理する。
AP層①では、言語毎に役割を分割する。
アクセス層①以降の翻訳エンジン部分については、言語及び書類公報毎に分割した構成とする。
また、審査官向OPDの翻訳エンジン部分については、より高い性能が求められるためさらに分離した構成とする。

1.2 システム間連携

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

案1では、以下に示す方式で要求元システムと通信を行なう。

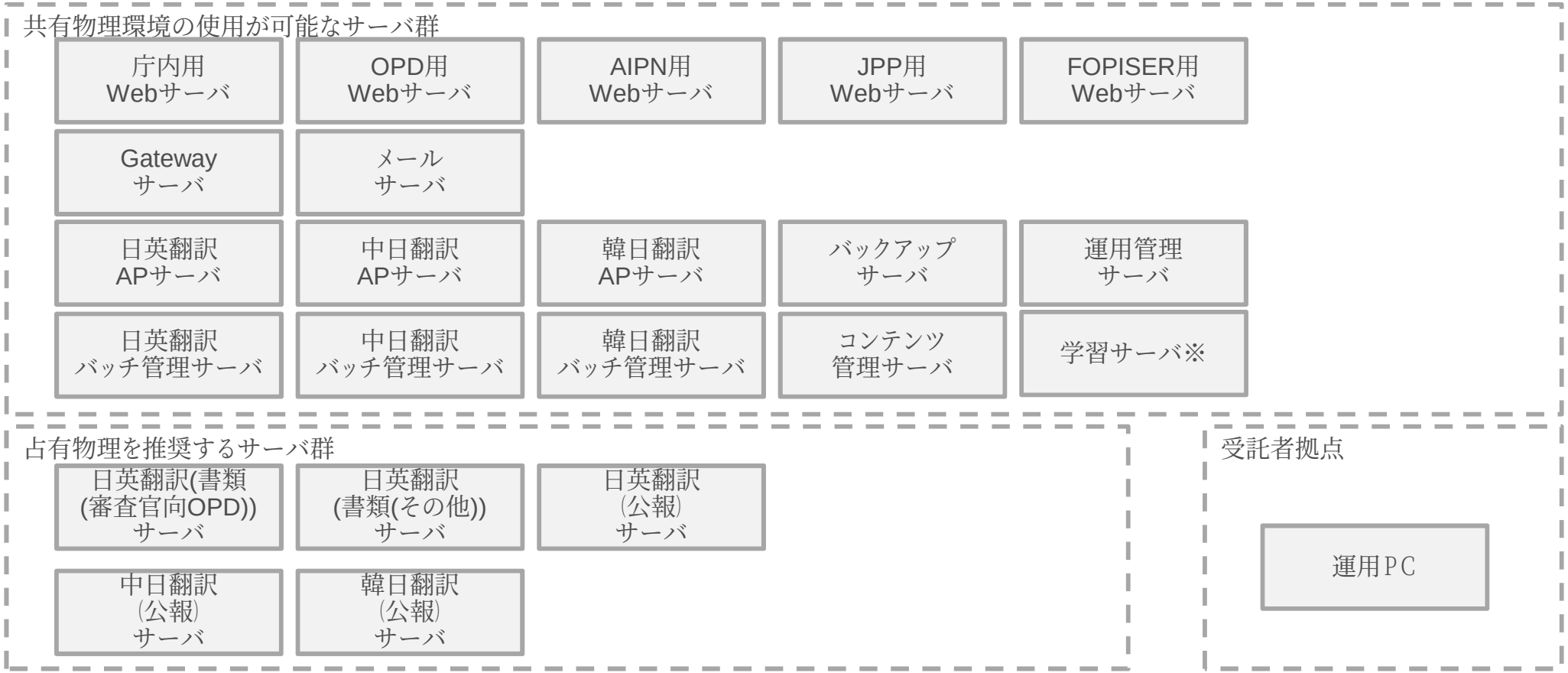


2 システム構成

2.1 機械翻訳に係るシステム全体構成図

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

案1のシステム全体構成図は以下の通り。
可能な限り費用の安い共有物理環境上にサーバを配置する。
ただし、共有物理環境は他サイトの使用状況により処理性能へ影響が出る場合があるため、
もっとも性能が求められる翻訳サーバについては、占有物理環境に配置する。

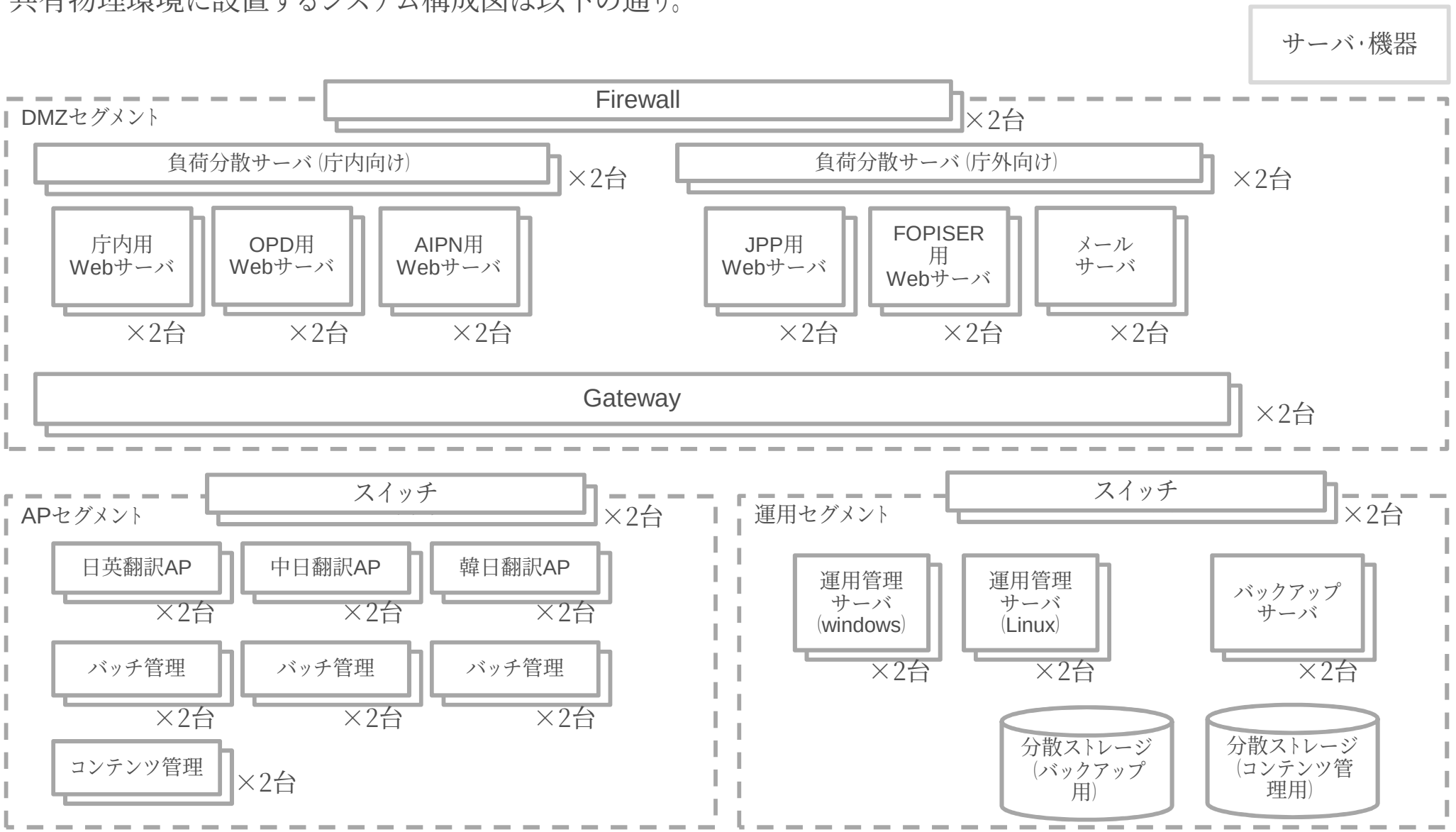


※学習サーバについては、性能上共有物理環境への設置が可能だが、必要なメモリ量が2.5TBを超えている。
これは、共有物理環境にて用意可能な容量（最大1.9TB）を超えているため、占有物理前提で見積を実施する。

2.2 機械翻訳システム構成図

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

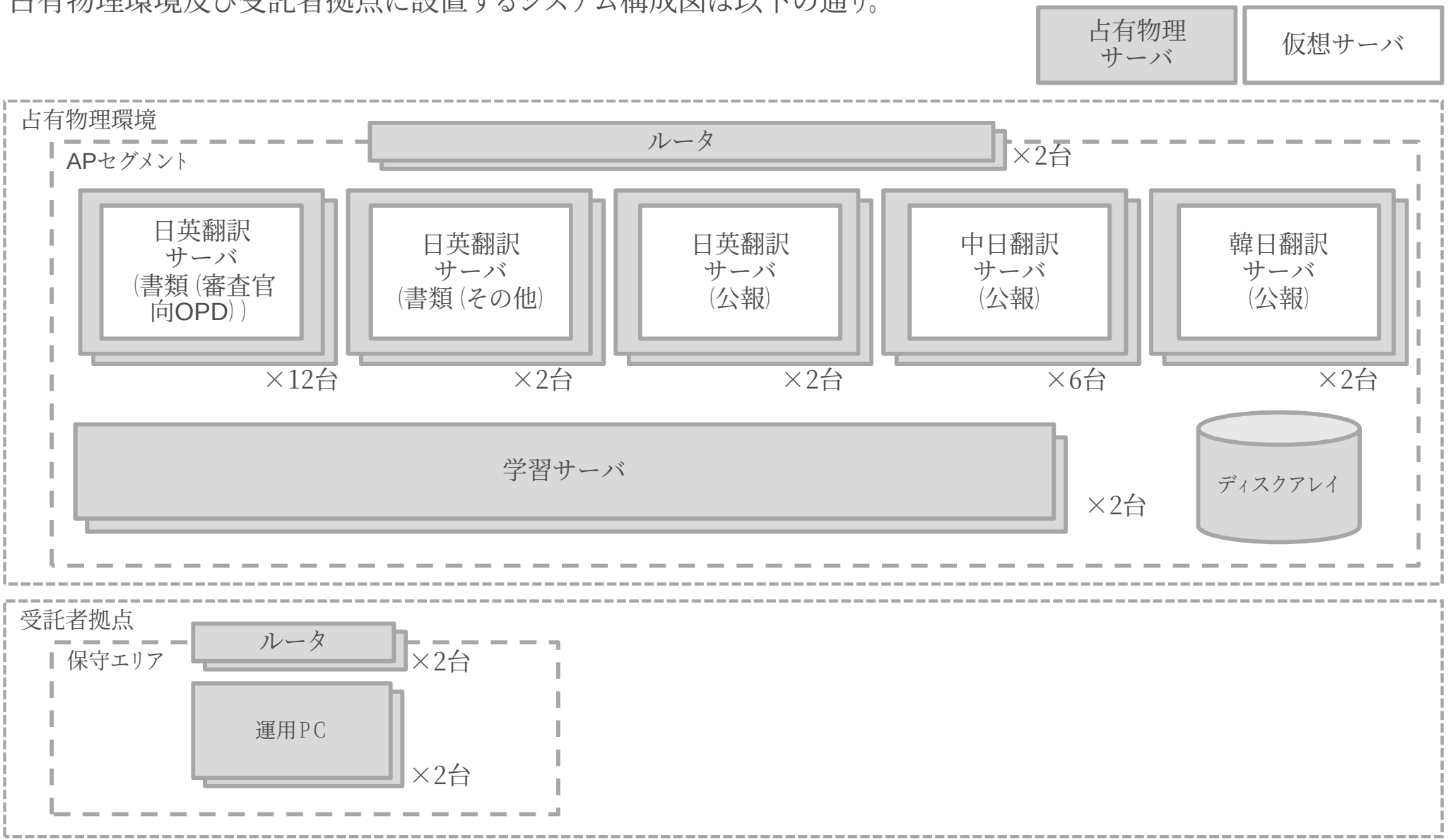
共有物理環境に設置するシステム構成図は以下の通り。



2.2 機械翻訳システム構成図

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

占有物理環境及び受託者拠点に設置するシステム構成図は以下の通り。



2.2 機械翻訳システム構成図

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

案1にて使用するハードウェアの一覧は以下の通り。

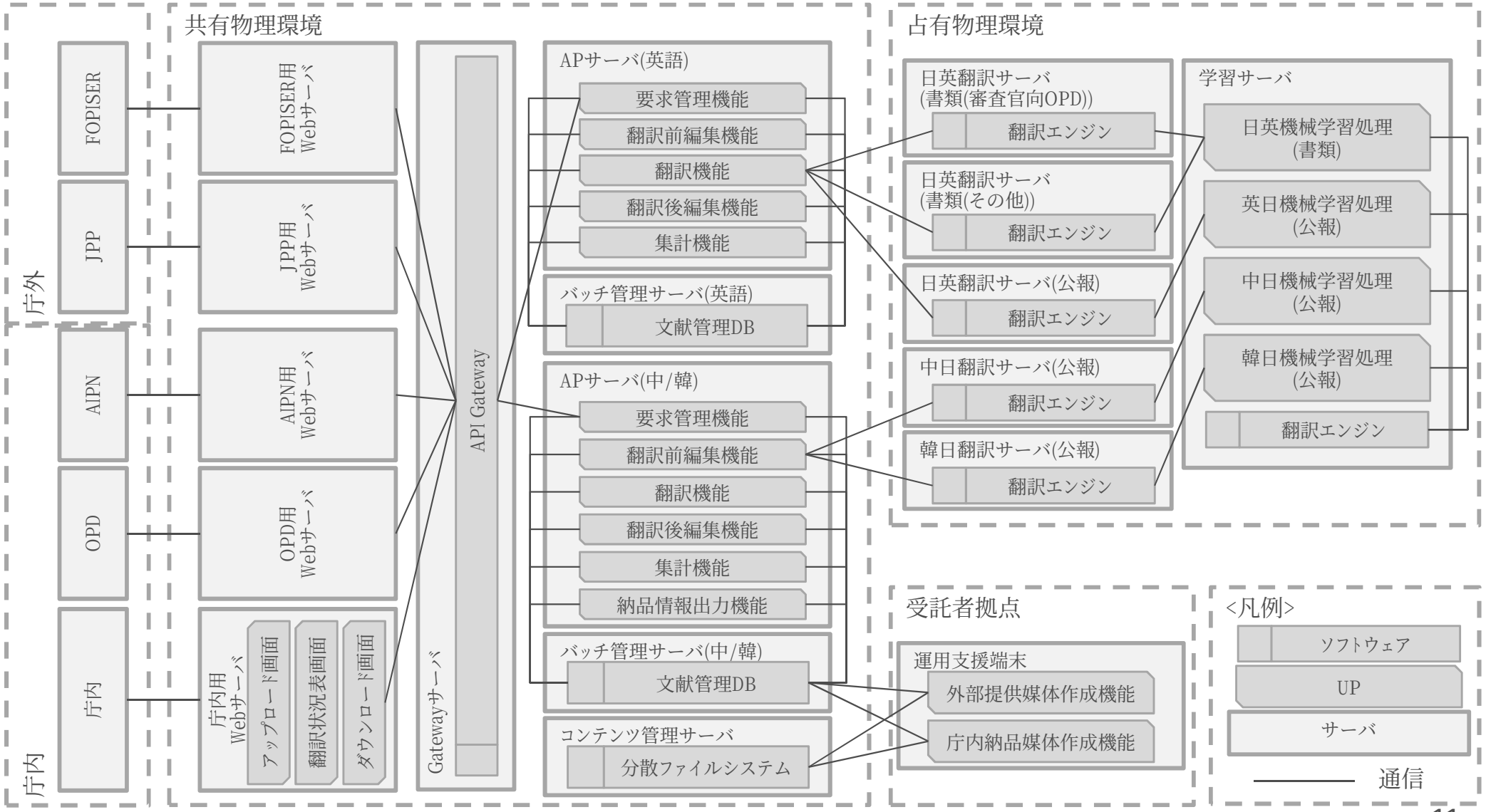
#	サーバ名	物理サーバ台数	仮想サーバ台数	サーバ台数小計
1	firewall用サーバ	0	2	2
2	負荷分散サーバ	0	4	4
3	Webサーバ	0	10	10
4	メールサーバ	0	2	2
5	Gatewayサーバ	0	2	2
6	APサーバ	0	6	6
7	バッチ管理サーバ	0	6	6
8	コンテンツ管理サーバ	0	2	2
9	運用管理サーバ(Linux)	0	2	2
10	運用管理サーバ(windows)	0	2	2
11	バックアップサーバ	0	2	2
12	翻訳サーバ	24	43	67
13	学習サーバ	2	0	2
14	運用PC	2	0	2
合計		28	83	111

3 システム方式設計

3.1 システム処理方式

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

案1では、以下に示すシステム処理方式を採用する。各機能の概要については、次頁参照。



3.2 機能

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

案1にて作成するプログラムの機能名と機能概要、各言語ごとの対応を以下に示す。

#	機能名	機能概要	英語 (書類)		英語 (公報)		中国 (公報)	韓国 (公報)
			英→日	日→英	英→日	日→英	中→日	韓→日
1	要求管理機能	翻訳要求を受信し各種機能の呼び出し制御を行ない、要求元システムに応答を返す機能	○				○	○
2	翻訳前編集機能	受信した文章を、翻訳処理ができる形式に編集する機能	○	○	○	○	○	○
3	翻訳機能	文献を文単位に分割し、翻訳サーバにて文章の翻訳を実施する機能	○	○	○	○	○	○
4	翻訳後編集機能	翻訳した文章を、受信時と同じ形式に編集し、要求元へ送信する機能	○	○	○	○	○	○
5	集計機能	翻訳時に発生した未知語や、運用保守に係る統計情報を集計する機能	○				○	○
6	文献アップロード機能 (画面)	特許庁運用担当者が翻訳対象の文献をアップロードする機能	ー	ー	ー	ー	○	○
7	文献ステータス表示機能 (画面)	翻訳中の文献について、処理状況を表示する機能	ー	ー	ー	ー	○	○
8	文献ダウンロード機能 (画面)	特許庁運用担当者が翻訳完了データやエラーリスト等をダウンロードする機能	ー	ー	ー	ー	○	○
9	納品情報出力機能	各種納品データをそれぞれの納品形式に編集して出力する機能	ー	ー	ー	ー	○	○

3.3 性能･拡張性

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

性能に関する拡張性については、各構成案間での差分は無く、各処理のプロセスの増設にて対応を行う。
詳細な拡張例については、要件整理資料の作成時に整理を実施する。

案1にて機能拡張する場合の拡張パターンと拡張方法は以下の通り。

#	機能名	種別追加時		言語追加時		言語方向追加時	
		改造	新規作成	改造	新規作成	改造	新規作成
1	要求管理機能	○	―	―	○	○	―
2	翻訳前編集機能	―	○	―	○	―	○
3	翻訳機能	―	○	―	○	―	○
4	翻訳後編集機能	―	○	―	○	―	○
5	集計機能	○	―	○	―	○	―
6	文献アップロード機能(画面)	○	―	○	―	○	―
7	文献ステータス表示機能(画面)	○	―	○	―	○	―
8	文献ダウンロード機能(画面)	○	―	○	―	○	―
9	納品情報出力機能	○	―	○	―	○	―

3.4 信頼性

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

本項目については、前提事項であり、実現方式が複数存在しないため
別途要件整理時に整理を実施する

3.5 運用

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

本項目については、前提事項であり、実現方式が複数存在しないため
別途要件整理時に整理を実施する

3.6 セキュリティ

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

本項目については、前提事項であり、実現方式が複数存在しないため
別途要件整理時に整理を実施する

4 コスト

4.1 コスト

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

案1の一時経費、運用費トータルコストの概算見積値との差分は以下の通り。

#	区分	項目	想定 との比率	想定との差異理由
1	一時経費	ハードウェア(導入経費含む)	0.62	一部サーバを共有物理環境上に構築することで物理サーバが減少したため
2		ソフトウェア	1.09	サーバ台数増加及び仮想ホスト分のソフトウェア追加のため
3		DC	1.00	—
4		ネットワーク	1.00	—
5		プログラム開発	1.00	—
6		環境構築	1.15	サーバ台数増加及び仮想ホスト分の構築作業追加のため
7	一時経費全体比率		0.82	—
8	運用費	ハードウェア保守(共有物理含む)	0.89	一部サーバを共有物理環境上に構築することで物理サーバが減少したため
9		ソフトウェア保守	1.09	サーバ台数増加及び仮想ホスト分のソフトウェア追加のため
10		DC	0.35	一部サーバを共有物理環境上に構築することで物理サーバが減少したため
11		ネットワーク	1.00	—
12		稼働維持(訳質向上含む)	1.15	管理対象のサーバが増えたため
13		蓄積等運用費	0.43	文献ダウンロード機能により、文献データ取得作業が不要になったため。
14	運用費全体比率		0.80	—
15	トータルコスト比率		0.81	—

5 機能適用技術

5.1 適用ソフトウェアの要件

案1	言語毎にアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

案1にて必要なソフトウェアの要件は以下の通り。

#	区分	機能要件
1	OS	-
2	WEB	Webサーバとしての機能を有し、HTTP/HTTPSによる通信が可能なこと
3		複数のサービスにて使用しているAPI(インターフェース)を統合管理する機能を有すること
4	DB	・RDBMSとしての機能を有すること ・他のソフトウェアとの連携又は自らの機能にて可用性向上を行なえること
5	機械翻訳ソフト	・対訳コーパスをインプットとして機械学習を行い翻訳モデルを作成できること ・作成した翻訳モデルを使用して機械翻訳が可能であること ・未知語等については、対訳辞書等を使用して訳質を向上させるためのチューニングができること
6	開発言語	機能一覧に記載した機能実現可能なこと
7	AP	作成したプログラムを実行可能なこと
8	分散ﾌｧｲﾙｼｽﾃﾑ	複数のホストがコンピュータネットワークを経由して共有しつつファイルにアクセスできること
9	分散処理基盤	複数のリソースを使用して与えられたデータセットを並列処理で処理する機能を有すること
10	ﾌｧｲﾙ操作	仮想管理基盤等と連携してファイルの操作等が可能なこと
11	バックアップ	クライアント/サーバ型のバックアップシステムを構築可能であること
12	ジョブ管理	自動でジョブを実行する機能を有し実行状況等の管理が可能であること
13	監視	様々なネットワークサービス、サーバ 、その他のネットワークハードウェアのステータスを監視･追跡する機能を有すること
14	ウィルス対策	・ウィルスを検知し、検疫する機能を有すること ・パターンファイルが定期的に更新されていること ・パターンファイルの更新が可能なこと
15	クラスタ	サーバやソフトウェア、プロセス等に対して信頼性向上のための機能を有すること

END

将来の機械翻訳システム構成案　案2

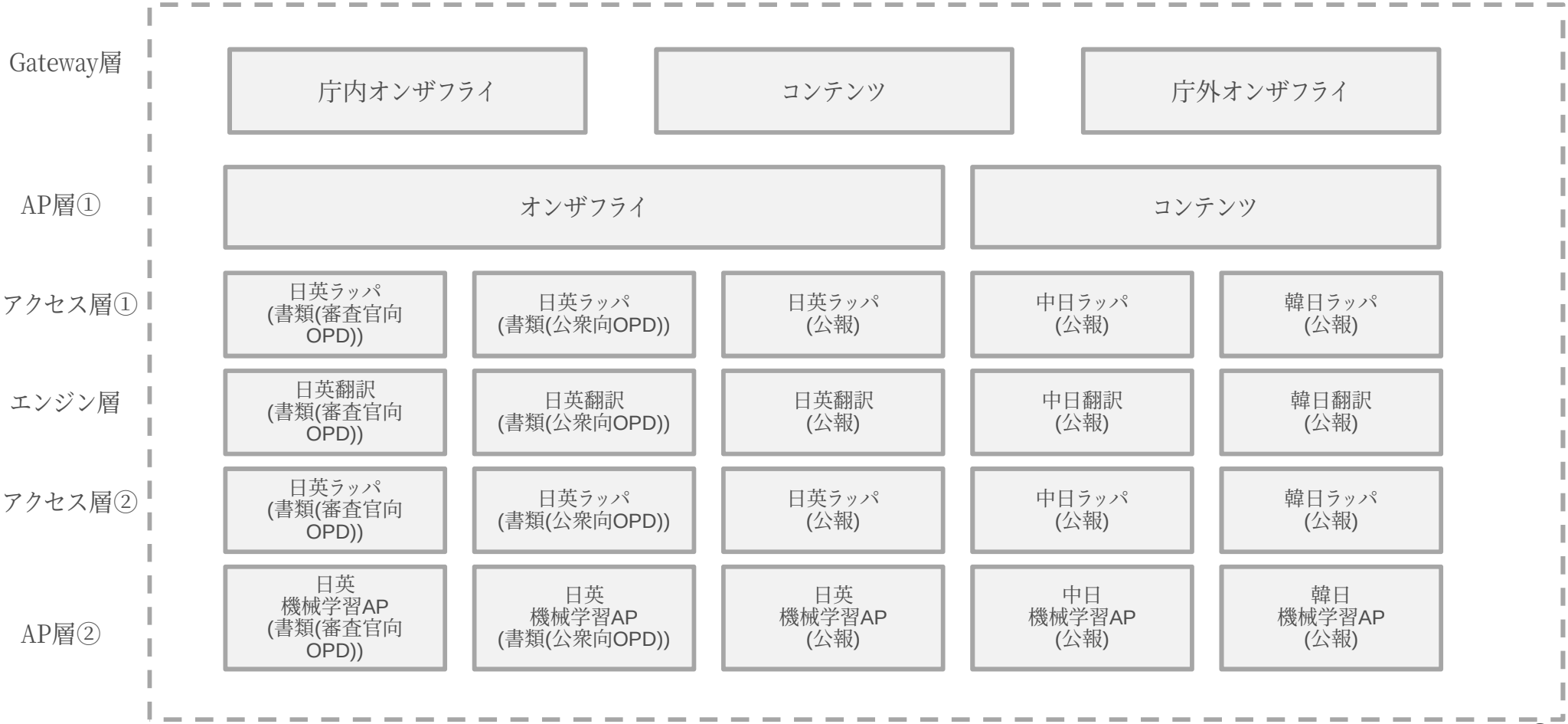
構成案　案2　目次

1 システムアーキテクチャ方針

1.1 アーキテクチャ概念図

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

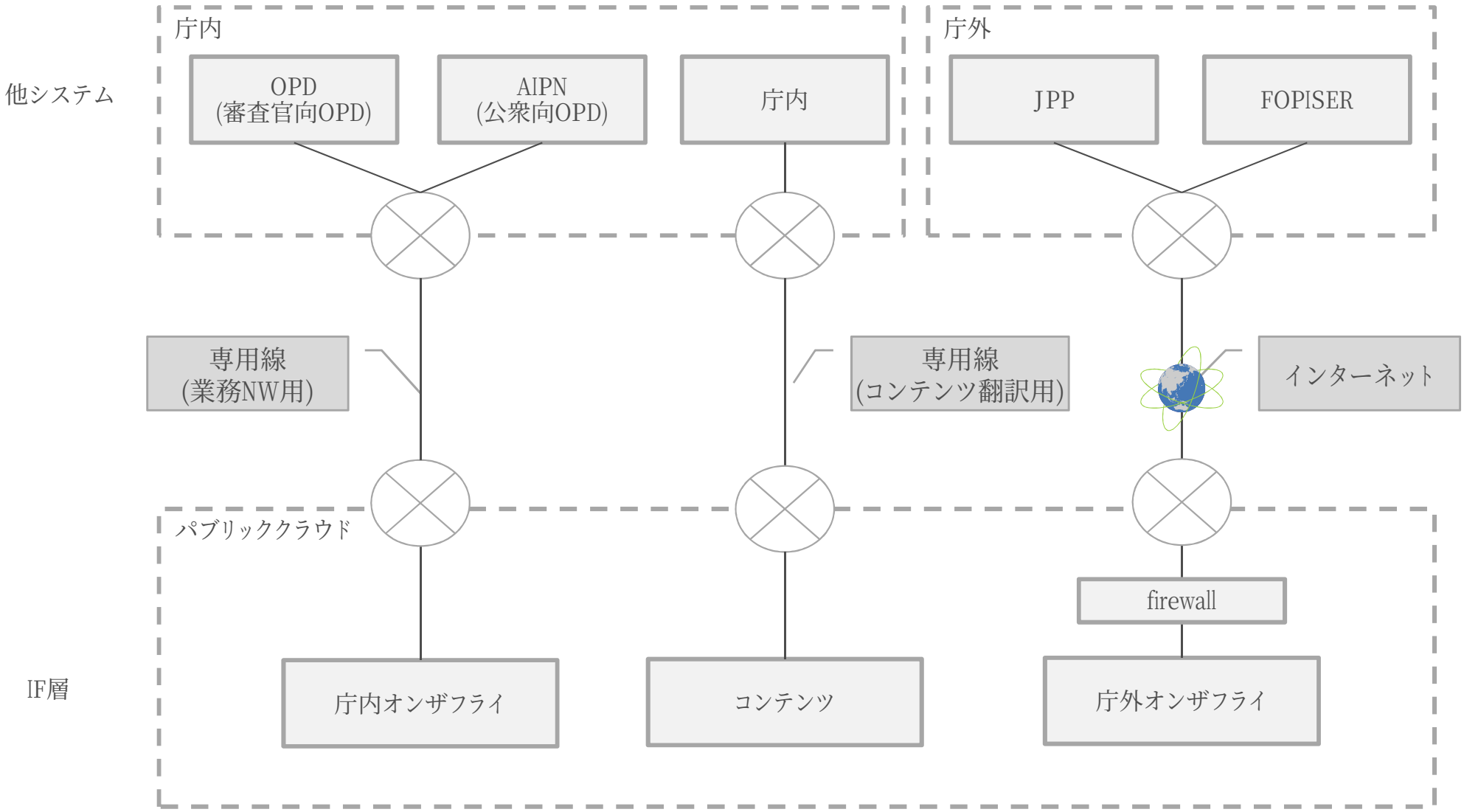
案2にて採用するアーキテクチャは以下の通り。
案2では、IF層をオンザフライ(庁内/庁外)とコンテンツの要求別に分割する。
AP層①では、オンザフライ、コンテンツの要求別に分割する。
アクセス層①以降の翻訳エンジン部分については、言語及び書類公報毎に分割した構成とする。
また、IF層、審査官向OPDの翻訳エンジン部分については、より高い性能が求められるためさらに分離した構成とする。



1.2 システム間連携

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

案2では、要求元システムとの連携を以下に示す方式で実施する

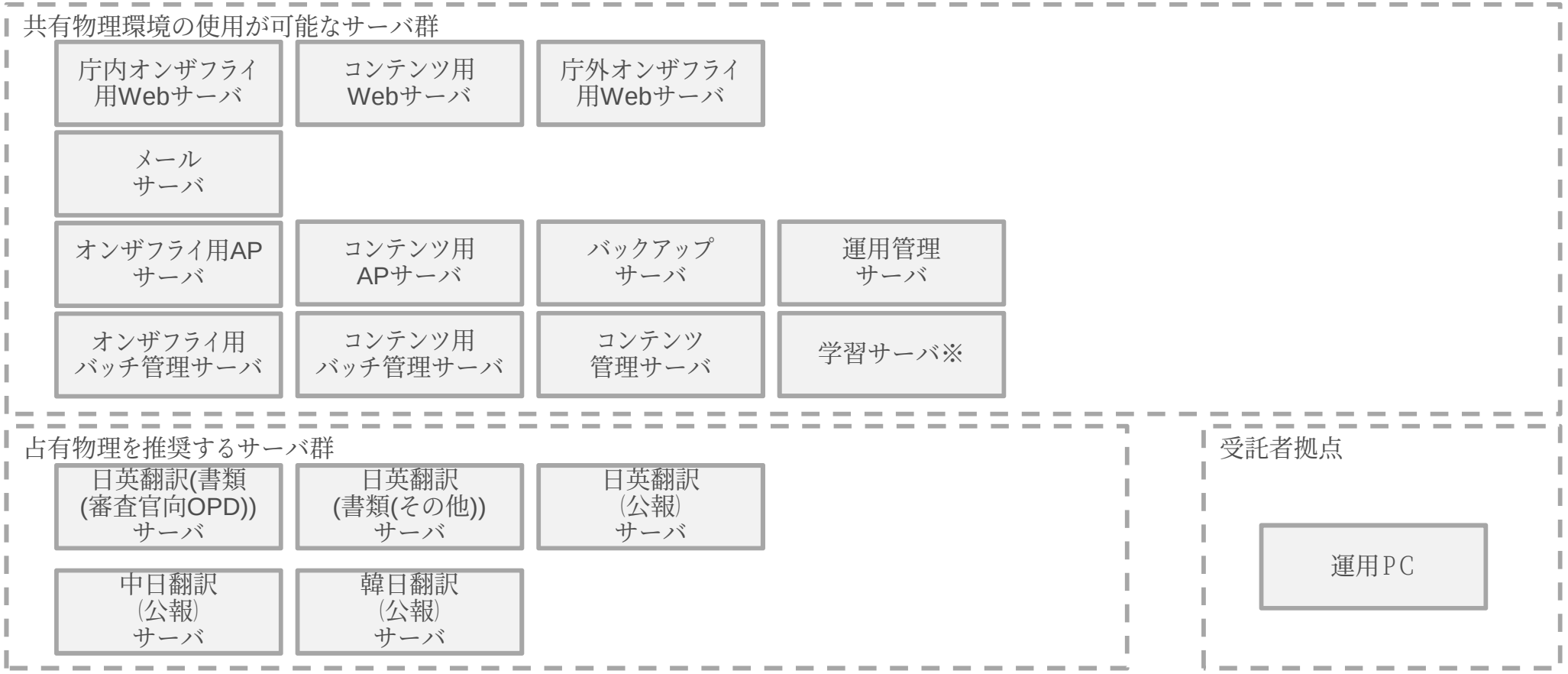


2 システム構成

2.1 機械翻訳に係るシステム全体構成図

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

案2のシステム全体構成図は以下の通り。
可能な限り費用の安い共有物理環境上にサーバを配置する。
ただし、共有物理環境は他サイトの使用状況により処理性能へ影響が出る場合があるため、
もっとも性能が求められる翻訳サーバについては、占有物理環境に配置する。

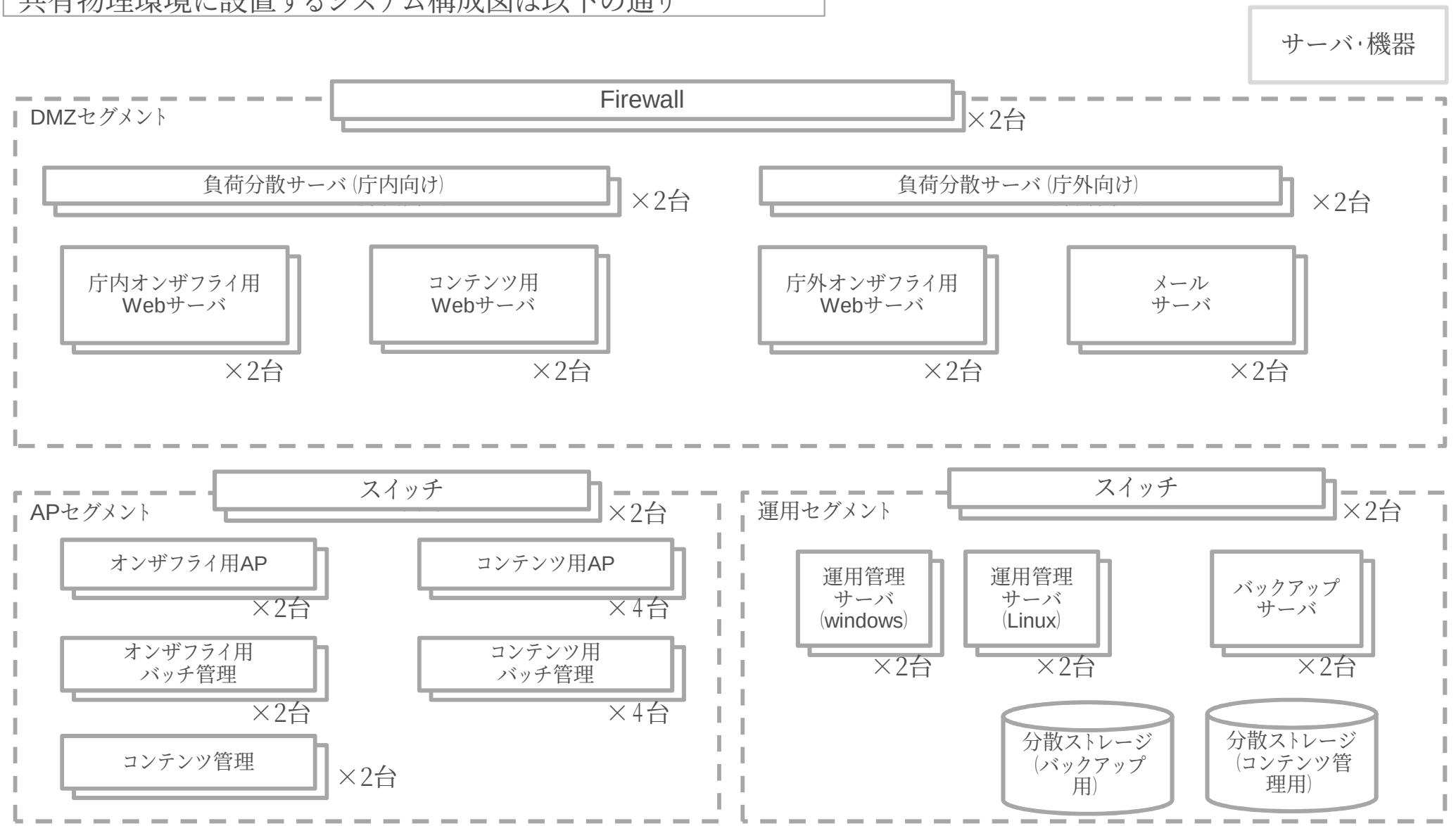


※学習サーバについては、性能上共有物理環境への設置が可能だが、必要なメモリ量が2.5TBを超えている。
これは、共有物理環境にて用意可能な容量（最大1.9TB）を超えているため、占有物理前提で見積を実施する。

2.2 機械翻訳システム構成図

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

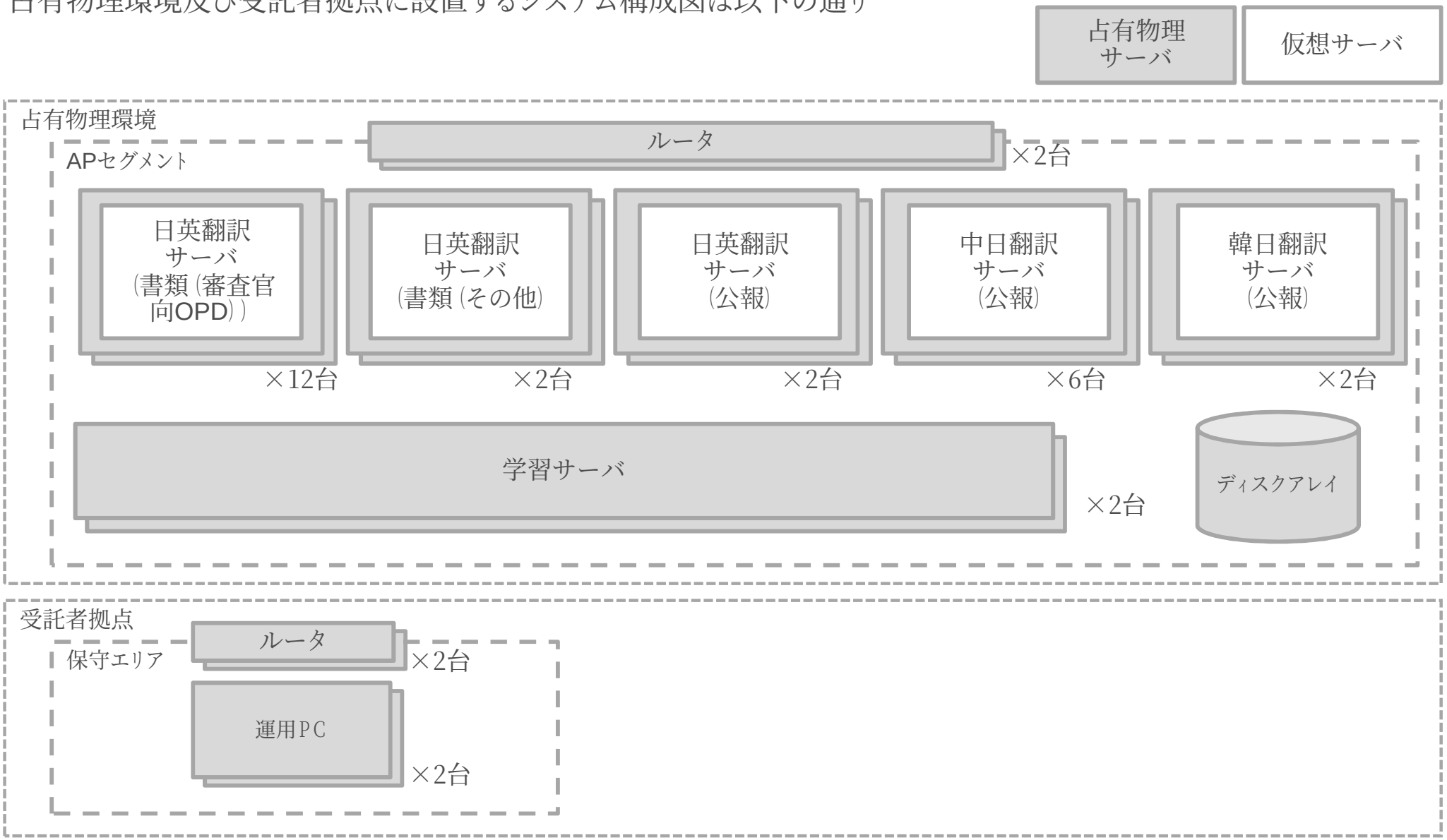
共有物理環境に設置するシステム構成図は以下の通り



2.2 機械翻訳システム構成図

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

占有物理環境及び受託者拠点に設置するシステム構成図は以下の通り



2.2 機械翻訳システム構成図

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

案2にて使用するハードウェアの一覧は以下の通り。

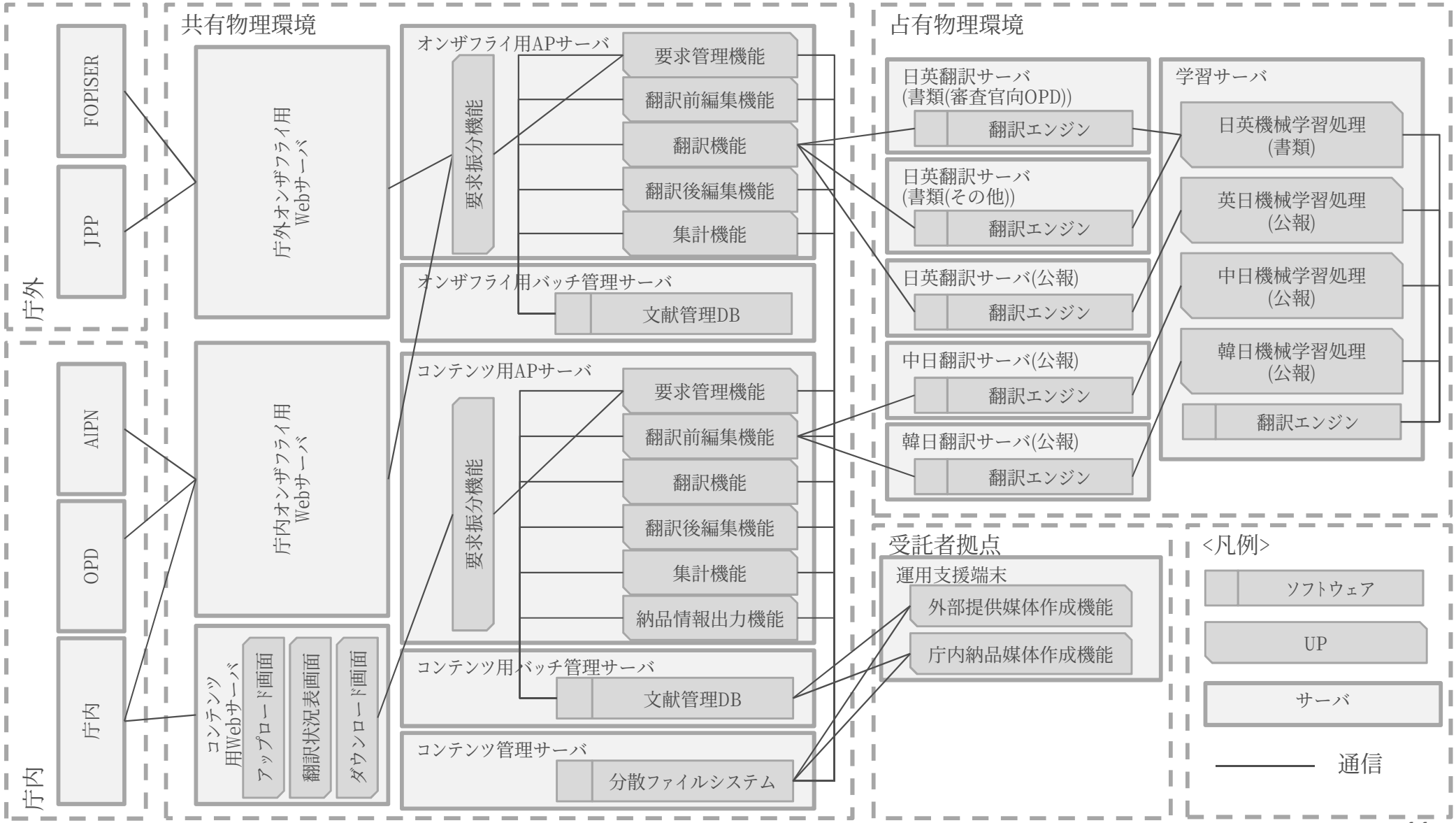
#	サーバ名	物理サーバ台数	仮想サーバ台数	サーバ台数小計
1	firewall用サーバ	0	2	2
2	負荷分散サーバ	0	4	4
3	Webサーバ	0	6	6
4	メールサーバ	0	2	2
5	APサーバ	0	6	6
6	バッチ管理サーバ	0	6	6
7	コンテンツ管理サーバ	0	2	2
8	運用管理サーバ(Linux)	0	2	2
9	運用管理サーバ(windows)	0	2	2
10	バックアップサーバ	0	2	2
11	翻訳サーバ	24	43	67
12	学習サーバ	2	0	2
13	運用PC	2	0	2
合計		28	77	105

3 システム方式設計

3.1 システム処理方式

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

案2では、以下に示すシステム処理方式を採用する。各機能の概要については、次頁参照。



3.2 機能

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

案2にて作成するプログラムの機能名と機能概要、各言語ごとの対応を以下に示す。

#	機能名	機能概要	オンザフライ				コンテンツ	
			英語(書類)		英語(公報)		中国 (公報)	韓国(公報)
			英→日	日→英	英→日	日→英	中→日	韓→日
1	要求振分機能	受信した要求から処理を行うサービスを決定し、要求元システムに処理結果を返す機能	○				○	
2	要求管理機能	各種機能の呼び出し制御を行う機能	○				○	
3	翻訳前編集機能	受信した文章を、翻訳処理ができる形式に編集する機能	○	○	○	○	○	○
4	翻訳機能	文献を文単位に分割し、翻訳サーバにて文章の翻訳を実施する機能	○	○	○	○	○	○
5	翻訳後編集機能	翻訳した文章を、受信時と同じ形式に編集し、要求元へ送信する機能	○	○	○	○	○	○
6	集計機能	翻訳時に発生した未知語や、運用保守に係る統計情報を集計する機能	○				○	
7	文献アップロード機能(画面)	特許庁運用担当者が翻訳対象の文献をアップロードする機能	ー	ー	ー	ー	○	○
8	文献ステータス表示機能(画面)	翻訳中の文献について、処理状況を表示する機能	ー	ー	ー	ー	○	○
9	文献ダウンロード機能(画面)	特許庁運用担当者が翻訳完了データやエラーリスト等をダウンロードする機能	ー	ー	ー	ー	○	○
10	納品情報出力機能	各種納品データをそれぞれの納品形式に編集して出力する機能	ー	ー	ー	ー	○	○

3.3 性能･拡張性

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

性能に関する拡張性については、各構成案間での差分は無く、各処理のプロセスの増設にて対応を行う。
詳細な拡張例については、要件整理資料の作成時に整理を実施する。

案2にて機能拡張する場合の拡張パターンと拡張方法は以下の通り。

#	機能名	種別追加時		言語追加時		言語方向追加時	
		改造	新規作成	改造	新規作成	改造	新規作成
1	要求振分機能	○	―	○	―	○	―
2	要求管理機能	○	―	○	―	○	―
3	翻訳前編集機能	―	○	―	○	―	○
4	翻訳機能	―	○	―	○	―	○
5	翻訳後編集機能	―	○	―	○	―	○
6	集計機能	○	―	○	―	○	―
7	文献アップロード機能(画面)	○	―	○	―	○	―
8	文献ステータス表示機能(画面)	○	―	○	―	○	―
9	文献ダウンロード機能(画面)	○	―	○	―	○	―
10	納品情報出力機能	○	―	○	―	○	―

3.4 信頼性

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

本項目については、前提事項であり、実現方式が複数存在しないため
別途要件整理時に整理を実施する

3.5 運用

案2	オンザフライ、コンテンツでアプリケーションを分割し、 言語及び書類 公報毎に翻訳機能を分割する

本項目については、前提事項であり、実現方式が複数存在しないため
別途要件整理時に整理を実施する

3.6 セキュリティ

案2	オンザフライ、コンテンツでアプリケーションを分割し、 言語及び書類、公報毎に翻訳機能を分割する

本項目については、前提事項であり、実現方式が複数存在しないため
別途要件整理時に整理を実施する

4 コスト

4.1 コスト

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

案2の一時経費、運用費トータルコストの概算見積値との差分は以下の通り。

#	区分	項目	想定との比率	想定との差異理由
1	一時経費	ハードウェア(導入経費含む)	0.60	一部サーバを共有物理環境上に構築することで物理サーバが減少したため
2		ソフトウェア	1.08	サーバ台数増加及び仮想ホスト分のソフトウェア追加のため
3		DC	1.00	—
4		ネットワーク	1.00	—
5		プログラム開発	1.32	要求振分機能を追加作成するため
6		環境構築	1.05	サーバ台数増加及び仮想ホスト分の構築作業追加のため
7	一時経費全体比率※		0.84	—
8	運用費	ハードウェア保守(共有物理含む)	0.82	一部サーバを共有物理環境上に構築することで物理サーバが減少したため
9		ソフトウェア保守	1.08	サーバ台数増加及び仮想ホスト分のソフトウェア追加のため
10		DC	0.35	一部サーバを共有物理環境上に構築することで物理サーバが減少したため
11		ネットワーク	1.00	—
12		稼働維持(訳質向上含む)	1.04	管理対象のサーバが増えたため
13		蓄積等運用費	0.43	文献ダウンロード機能により、文献データ取得作業が不要になったため。
14	運用費全体比率		0.77	—
15	トータルコスト比率		0.81	—

5 機能適用技術

5.1 適用ソフトウェアの要件

案2	オンザフライ、コンテンツでアプリケーションを分割し、言語及び書類、公報毎に翻訳機能を分割する

案2にて必要なソフトウェアの要件は以下の通り。

#	区分	機能要件
1	OS	-
2	WEB	Webサーバとしての機能を有し、HTTP/HTTPSによる通信が可能なこと
3	DB	・RDBMSとしての機能を有すること ・他のソフトウェアとの連携又は自らの機能にて可用性向上を行なえること
4	機械翻訳ソフト	・対訳コーパスをインプットとして機械学習を行い翻訳モデルを作成できること ・作成した翻訳モデルを使用して機械翻訳が可能であること ・未知語等については、対訳辞書等を使用して訳質を向上させるためのチューニングができること
5	開発言語	機能一覧に記載した機能実現可能なこと
6	AP	作成したプログラムを実行可能なこと
7	分散ﾌｧｲﾙｼｽﾃﾑ	複数のホストがコンピュータネットワークを経由して共有しつつファイルにアクセスできること
8	分散処理基盤	複数のリソースを使用して与えられたデータセットを並列処理で処理する機能を有すること
9	ﾌｧｲﾙ操作	仮想管理基盤等と連携してファイルの操作等が可能なこと
10	バックアップ	クライアント/サーバ型のバックアップシステムを構築可能であること
11	ジョブ管理	自動でジョブを実行する機能を有し実行状況等の管理が可能であること
12	監視	様々なネットワークサービス、サーバ 、その他のネットワークハードウェアのステータスを監視･追跡する機能を有すること
13	ウィルス対策	・ウィルスを検知し、検疫する機能を有すること ・パターンファイルが定期的に更新されていること ・パターンファイルの更新が可能なこと
14	クラスタ	サーバやソフトウェア、プロセス等に対して信頼性向上のための機能を有すること

END

将来の機械翻訳システム構成案　案3

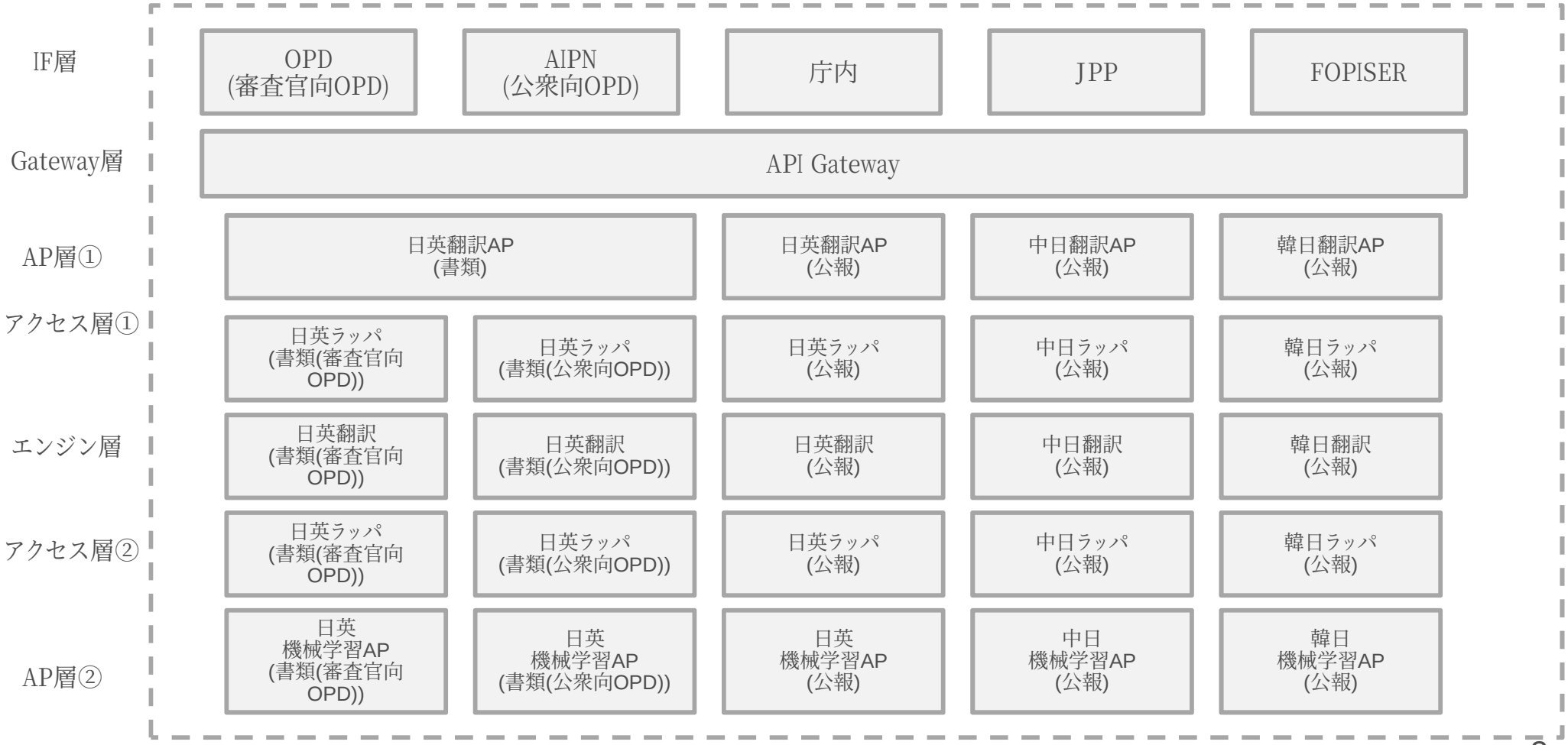
構成案　案3　目次

1 システムアーキテクチャ方針

1.1 アーキテクチャ概念図

案3	言語及び書類、公報毎に アプリケーションと翻訳機能を分割する

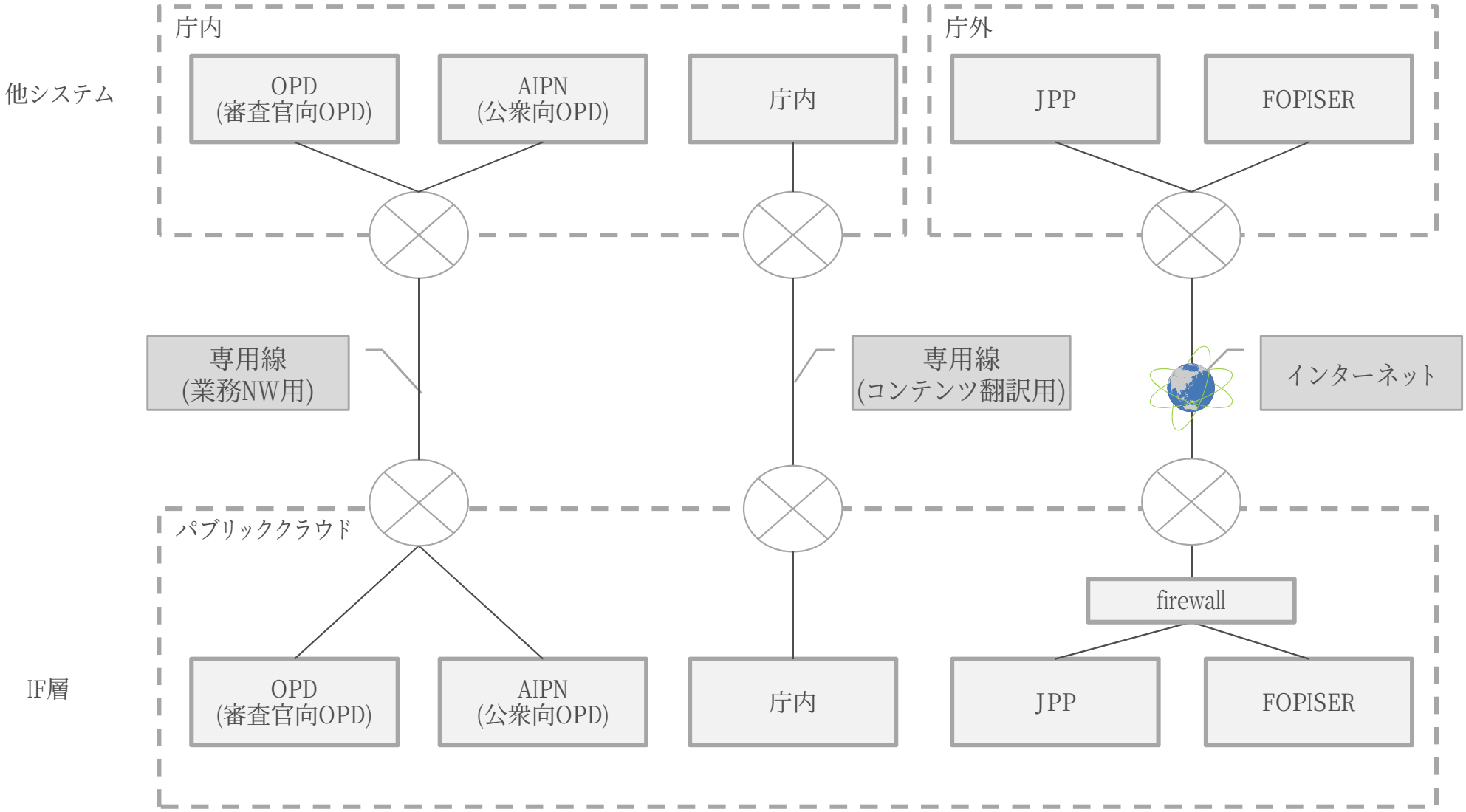
案3にて採用するアーキテクチャは以下の通り。
案3では、IF層を各要求元システム分用意し、各システムから来た要求をAPI Gatewayにて一括管理する。
AP層①では、言語及び書類、公報毎に役割を分割する。
アクセス層①以降の翻訳エンジン部分については、言語及び書類公報毎に分割した構成とする。
また、審査官向OPDの翻訳エンジン部分については、より高い性能が求められるためさらに分離した構成とする。



1.2 システム間連携

案3	言語及び書類 公報毎に アプリケーションと翻訳機能を分割する

案3では、要求元システムとの連携を以下に示す方式で実施する

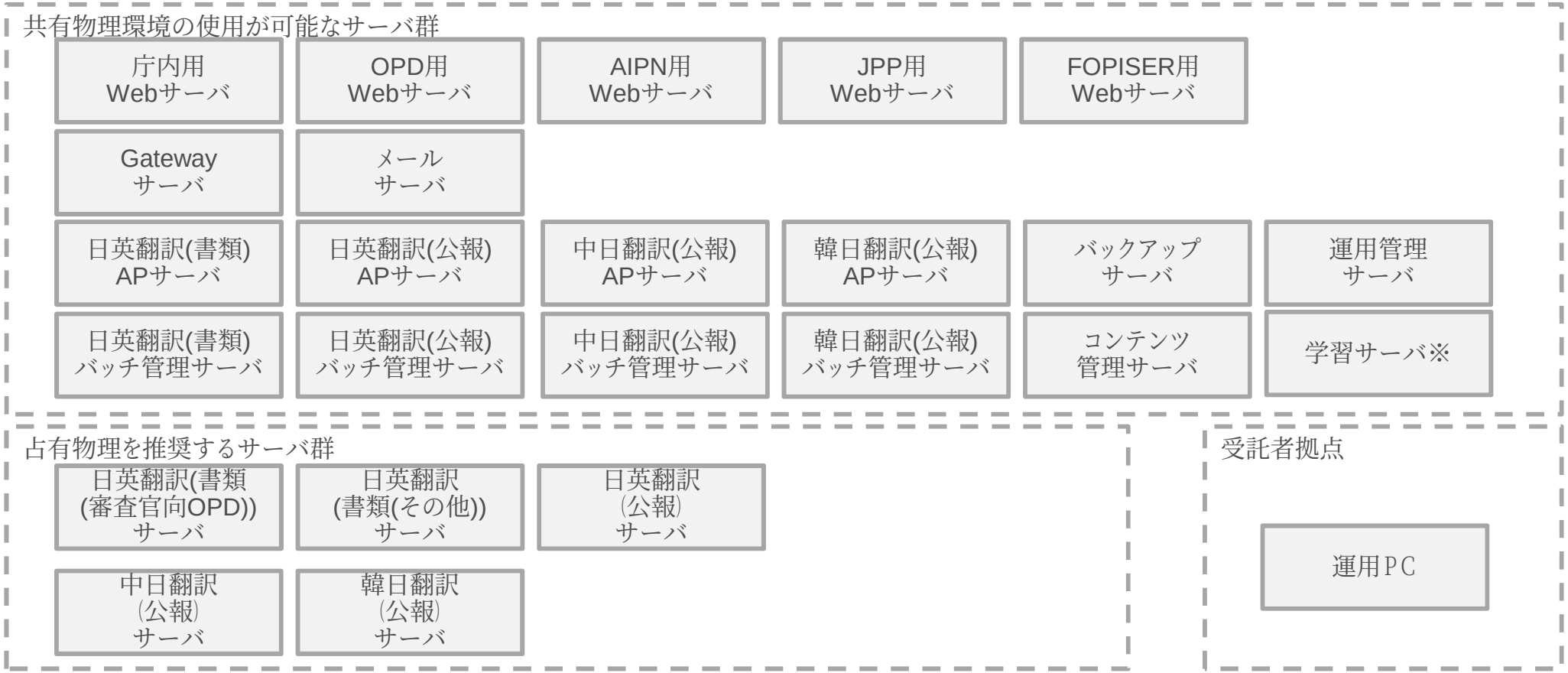


2 システム構成

2.1 機械翻訳に係るシステム全体構成図

案3	言語及び書類、公報毎に アプリケーションと翻訳機能を分割する

案3のシステム全体構成図は以下の通り。
可能な限り費用の安い共有物理環境上にサーバを配置する。
ただし、共有物理環境は他サイトの使用状況により処理性能へ影響が出る場合があるため、
もっとも性能が求められる翻訳サーバについては、占有物理環境に配置する。

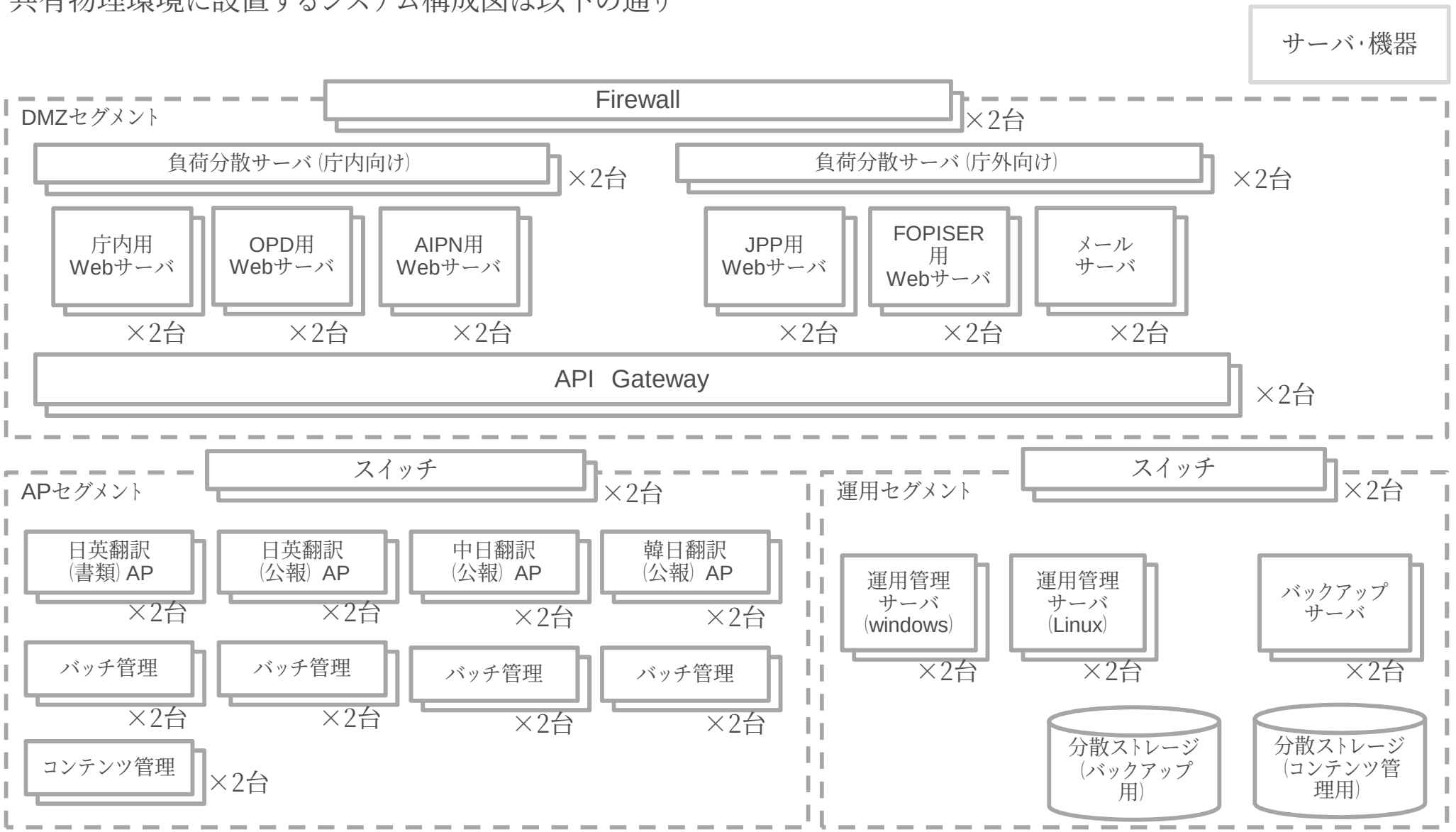


※学習サーバについては、性能上共有物理環境への設置が可能だが、必要なメモリ量が2.5TBを超えている。
これは、共有物理環境にて用意可能な容量（最大1.9TB）を超えているため、占有物理前提で見積を実施する。

2.2 機械翻訳システム構成図

案3	言語及び書類 公報毎に アプリケーションと翻訳機能を分割する

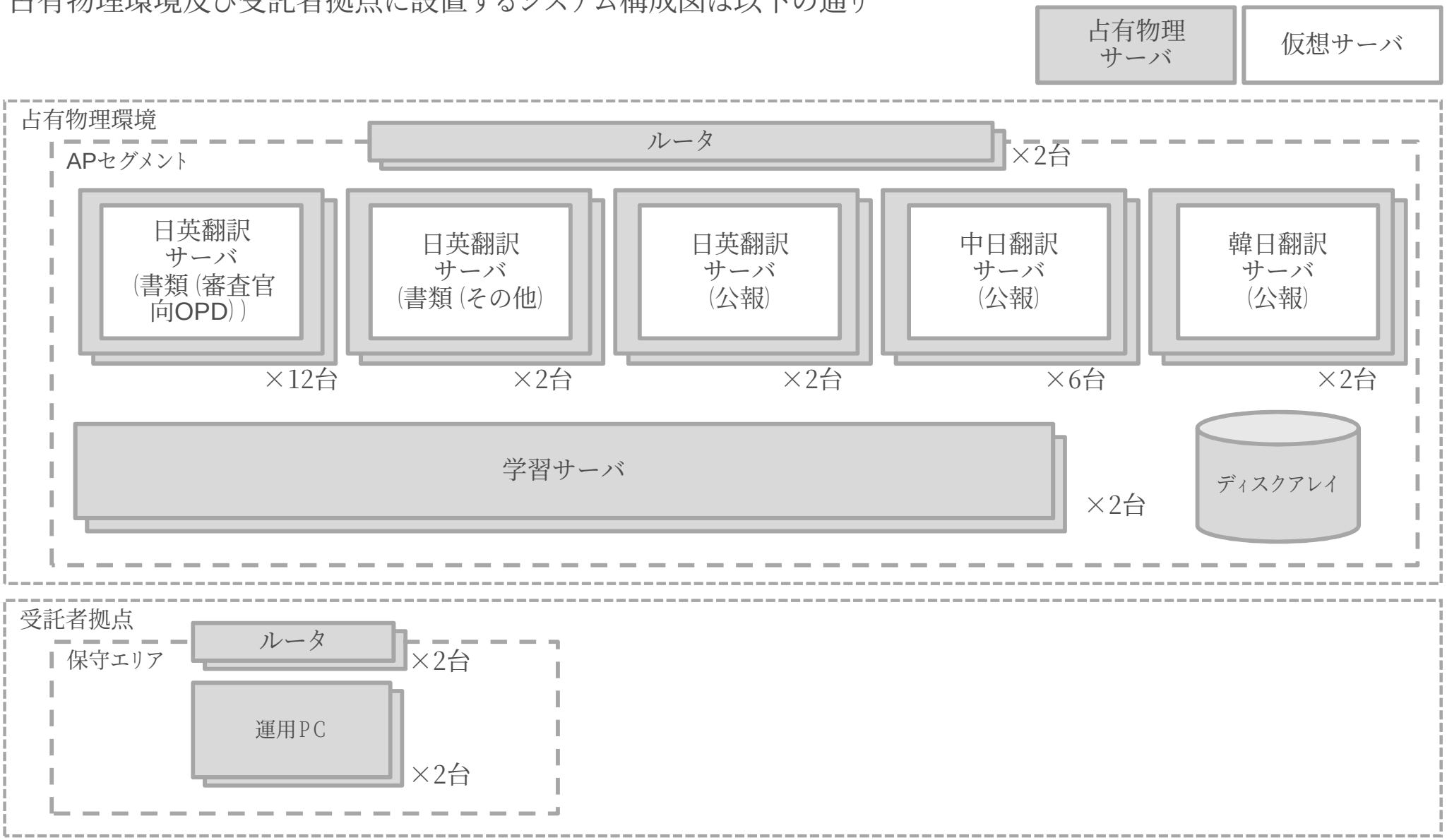
共有物理環境に設置するシステム構成図は以下の通り



2.2 機械翻訳システム構成図

案3	言語及び書類 公報毎に アプリケーションと翻訳機能を分割する

占有物理環境及び受託者拠点に設置するシステム構成図は以下の通り



2.2 機械翻訳システム構成図

案3	言語及び書類 公報毎に アプリケーションと翻訳機能を分割する

案3にて使用するハードウェアの一覧は以下の通り。

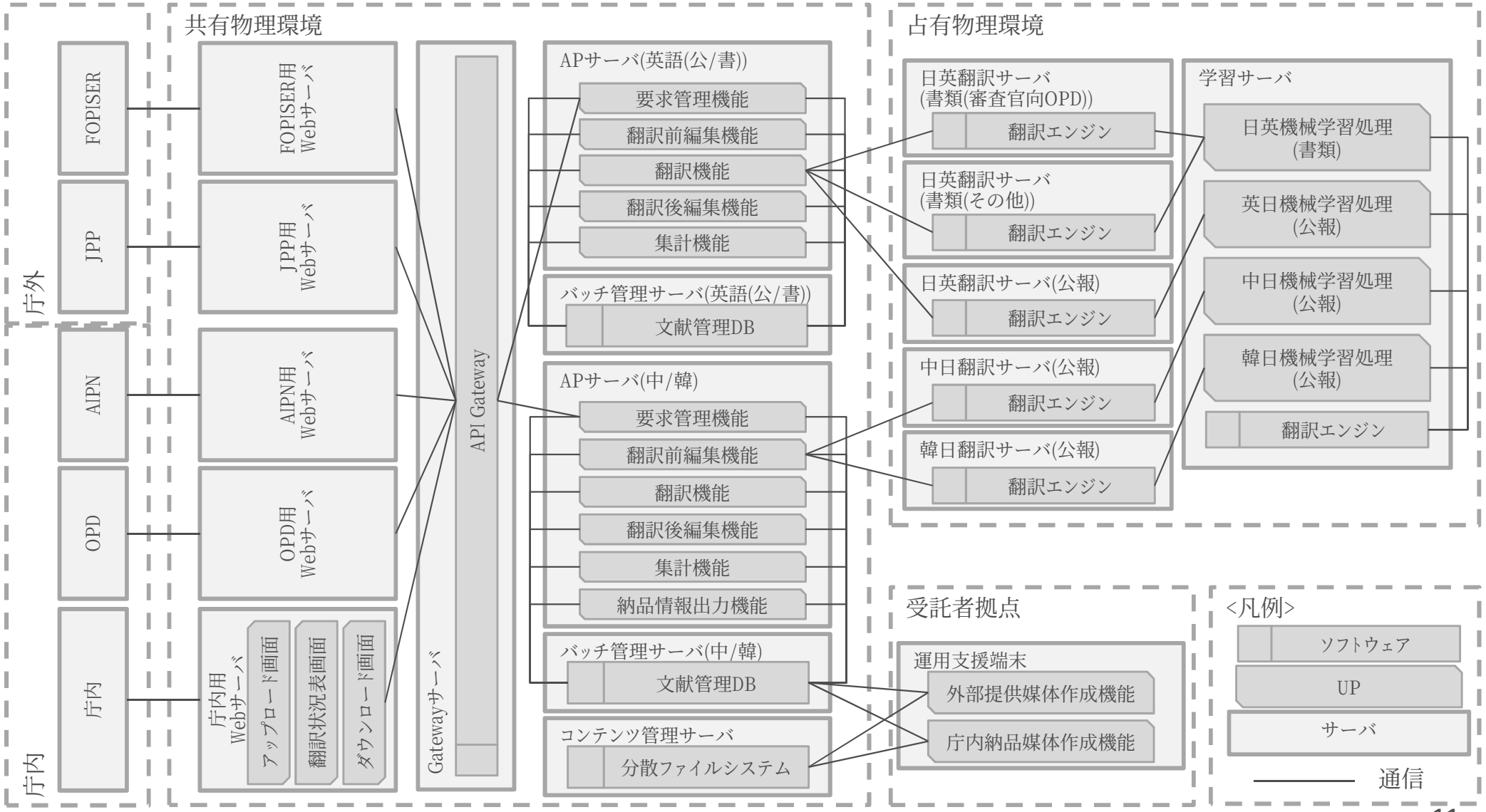
#	サーバ名	物理サーバ台数	仮想サーバ台数	サーバ台数小計
1	firewall用サーバ	0	2	2
2	負荷分散サーバ	0	4	4
3	Webサーバ	0	10	10
4	メールサーバ	0	2	2
5	Gatewayサーバ	0	2	2
6	APサーバ	0	8	8
7	バッチ管理サーバ	0	8	8
8	コンテンツ管理サーバ	0	2	2
9	運用管理サーバ (Linux)	0	2	2
10	運用管理サーバ (windows)	0	2	2
11	バックアップサーバ	0	2	2
12	翻訳サーバ	24	43	67
13	学習サーバ	2	0	2
14	運用PC	2	0	2
合計		28	83	111

3 システム方式設計

3.1 システム処理方式

案3	言語及び書類 公報毎に アプリケーションと翻訳機能を分割する

案3では、以下に示すシステム処理方式を採用する。各機能の概要については、次頁参照。



3.2 機能

案3	言語及び書類、公報毎に アプリケーションと翻訳機能を分割する

案3にて作成するプログラムの機能名と機能概要、各言語ごとの対応を以下に示す。

#	機能名	機能概要	英語 (書類)		英語 (公報)		中国 (公報)	韓国 (公報)
			英→日	日→英	英→日	日→英	中→日	韓→日
1	要求管理機能	翻訳要求を受信し各種機能の呼び出し制御を行ない、要求元システムに応答を返す機能	○		○		○	○
2	翻訳前編集機能	受信した文章を、翻訳処理ができる形式に編集する機能	○	○	○	○	○	○
3	翻訳機能	文献を文単位に分割し、翻訳サーバにて文章の翻訳を実施する機能	○	○	○	○	○	○
4	翻訳後編集機能	翻訳した文章を、受信時と同じ形式に編集し、要求元へ送信する機能	○	○	○	○	○	○
5	集計機能	翻訳時に発生した未知語や、運用保守に係る統計情報を集計する機能	○		○		○	○
6	文献アップロード機能 (画面)	特許庁運用担当者が翻訳対象の文献をアップロードする機能	ー	ー	ー	ー	○	○
7	文献ステータス表示機能 (画面)	翻訳中の文献について、処理状況を表示する機能	ー	ー	ー	ー	○	○
8	文献ダウンロード機能 (画面)	特許庁運用担当者が翻訳完了データやエラーリスト等をダウンロードする機能	ー	ー	ー	ー	○	○
9	納品情報出力機能	各種納品データをそれぞれの納品形式に編集して出力する機能	ー	ー	ー	ー	○	○

3.3 性能･拡張性

案3	言語及び書類 公報毎に アプリケーションと翻訳機能を分割する

性能に関する拡張性については、各構成案間での差分は無く、各処理のプロセスの増設にて対応を行う。
詳細な拡張例については、要件整理資料の作成時に整理を実施する。

案3にて機能拡張する場合の拡張パターンと拡張方法は以下の通り。

#	機能名	種別追加時		言語追加時		言語方向追加時	
		改造	新規作成	改造	新規作成	改造	新規作成
1	要求管理機能	―	○	―	○	○	―
2	翻訳前編集機能	―	○	―	○	―	○
3	翻訳機能	―	○	―	○	―	○
4	翻訳後編集機能	―	○	―	○	―	○
5	集計機能	―	○	―	○	○	―
6	文献アップロード機能(画面)	○	―	○	―	○	―
7	文献ステータス表示機能(画面)	○	―	○	―	○	―
8	文献ダウンロード機能(画面)	○	―	○	―	○	―
9	納品情報出力機能	○	―	○	―	○	―

3.4 信頼性

案3	言語及び書類 公報毎に アプリケーションと翻訳機能を分割する

構成案間での差分が無いため、
要件整理資料作成時に整理を実施

3.5 運用

案3	言語及び書類 公報毎に アプリケーションと翻訳機能を分割する

構成案間での差分が無いため、
要件整理資料作成時に整理を実施

3.6 セキュリティ

案3	言語及び書類 公報毎に アプリケーションと翻訳機能を分割する

構成案間での差分が無いため、
要件整理資料作成時に整理を実施

4 コスト

4.1 コスト

案3	言語及び書類 公報毎に アプリケーションと翻訳機能を分割する

案3の一時経費、運用費トータルコストの概算見積値との差分は以下の通り。

#	区分	項目	想定との比率	想定との差異理由
1	一時経費	ハードウェア(導入経費含む)	0.62	一部サーバを共有物理環境上に構築することで物理サーバが減少したため
2		ソフトウェア	1.10	サーバ台数増加及び仮想ホスト分のソフトウェア追加のため
3		DC	1.00	—
4		ネットワーク	1.00	—
5		プログラム開発	1.00	—
6		環境構築	1.20	サーバ台数増加及び仮想ホスト分の構築作業追加のため
7	一時経費全体比率※		0.83	—
8	運用費	ハードウェア保守(共有物理含む)	0.91	一部サーバを共有物理環境上に構築することで物理サーバが減少したため
9		ソフトウェア保守	1.09	サーバ台数増加及び仮想ホスト分のソフトウェア追加のため
10		DC	0.35	一部サーバを共有物理環境上に構築することで物理サーバが減少したため
11		ネットワーク	1.00	—
12		稼働維持(訳質向上含む)	1.20	管理対象のサーバが増えたため
13		蓄積等運用費	0.43	文献ダウンロード機能により、文献データ取得作業が不要になったため。
14	運用費全体比率		0.81	—
15	トータルコスト比率		0.82	—

5 機能適用技術

5.1 適用ソフトウェアの要件

案3	言語及び書類 公報毎に アプリケーションと翻訳機能を分割する

案3にて必要なソフトウェアの要件は以下の通り。

#	区分	機能要件
1	OS	-
2	WEB	Webサーバとしての機能を有し、HTTP/HTTPSによる通信が可能なこと
3		複数のサービスにて使用しているAPI(インターフェース)を統合管理する機能を有すること
4	DB	・RDBMSとしての機能を有すること ・他のソフトウェアとの連携又は自らの機能にて可用性向上を行なえること
5	機械翻訳ソフト	・対訳コーパスをインプットとして機械学習を行い翻訳モデルを作成できること ・作成した翻訳モデルを使用して機械翻訳が可能であること ・未知語等については、対訳辞書等を使用して訳質を向上させるためのチューニングができること
6	開発言語	機能一覧に記載した機能実現可能なこと
7	AP	作成したプログラムを実行可能なこと
8	分散ﾌｧｲﾙｼｽﾃﾑ	複数のホストがコンピュータネットワークを経由して共有しつつファイルにアクセスできること
9	分散処理基盤	複数のリソースを使用して与えられたデータセットを並列処理で処理する機能を有すること
10	ﾌｧｲﾙ操作	仮想管理基盤等と連携してファイルの操作等が可能なこと
11	バックアップ	クライアント/サーバ型のバックアップシステムを構築可能であること
12	ジョブ管理	自動でジョブを実行する機能を有し実行状況等の管理が可能であること
13	監視	様々なネットワークサービス、サーバ 、その他のネットワークハードウェアのステータスを監視･追跡する機能を有すること
14	ウィルス対策	・ウィルスを検知し、検疫する機能を有すること ・パターンファイルが定期的に更新されていること ・パターンファイルの更新が可能なこと
15	クラスタ	サーバやソフトウェア、プロセス等に対して信頼性向上のための機能を有すること

END