クラウドサーバリソース最適化による快適なWeb会議サービスの実現

本稿では，Mintent構成技術の1つである，「クラウドサーバリソース最適化技術」のWeb会議サービス適用事例を紹介します．本技術は，「ユーザ体感品質（QoE）を確保しつつ，リソースコストを削減したい」といったサービス提供者や，ユーザそれぞれの要件（インテント）を実現するために，さまざまなサービス利用状況に応じたパフォーマンスを予測するAI（人工知能）を用いて，品質確保に必要な最低限のリソース量にシステムを制御します．これにより，ユーザのサービス対するエンゲージメントの向上やサービス提供者の運用コストおよびリソースコストの削減，さらには，エネルギー使用効率化によるCO_2削減など，さまざまな課題の解決を実現します．

菊島（きくしま）　宏明（ひろあき）　呉（ご）　超（ちょう）
NTTアクセスサービスシステム研究所

はじめに

近年，リモートワークの急増やデジタルトランスフォーメーション（DX）推進など，予測不能な社会情勢の変化により，サービスに対する需要の急変やユーザニーズの多様化が加速しています．

それら環境の変化や多様性に追従する手段として各種システムのクラウドネイティブ化が進み，クラウドサーバ上でさまざまなサービスやネットワーク機能が迅速かつ多様な粒度で提供されるようになってきました．そのため，サーバリソース制御の重要性が増しており，これまでの経験頼りのリソース設計やシステムパフォーマンスベースでのリソース制御では，ユーザ品質要件や利用状況に則したサービス提供が困難となっています．そこでNTTアクセスサービスシステム研究所（AS研）では，システム視点に加えて，多様なユーザ要件を考慮した新たなリソース制御技術の確立を進めています．

新たなリソース制御技術では，ユーザ体感品質（QoE：Quality of Experience）を維持しつつ，さまざまなサービス利用状況下における最適リソースを自動算出することで，リソース設計や運用稼働の削減やリソースコストの削減につなげていきます．

また，本技術の有効性を確認するために，Web会議サービスに対して適応した事例を紹介し，Web会議に適用するために実施した新たな技術検討についても解説します．

クラウドサーバリソース最適化技術の概要

■技術の優位性

本技術は，システム負荷状況やサービス利用状況から，さまざまなパフォーマンスを予測するAI（人工知能）モデルを用いて，インテントを満たす最小リソース量を算出する技術です（**図1**）．

従来のシステムパフォーマンスのみに着目したリソース制御技術システムに比べて次のような優位性を持っています．

まず，本技術は各種パフォーマンスを予測する複数のAIモデルの組合せにより，システムパフォーマンス要件だけでなく，QoE要件も含めたインテントに対して最適なリソース量の算出が可能です．使用するAIモデルは，さまざまなログデータを活用して各種パフォーマンスを予測できるように学習されます．基本的には回帰分析モデルを使用しますが，QoE指標によっては，回帰分析による数値予測が困難な性質のものもあるため，その場合は，分類分析モデルによる品質レベルを予測するなど，使い方に応じて選択することが可能となっています（図1）．これらの特徴により，単なる品質故障の回避やリソース効率化といった観点だけでなく，ユーザの利用状況や目的に合わせたサービス提供が可能となります．

そのため，品質要件をより確実に満足できるようにするため，AIモデルの学習においても，さまざまな工夫がなされています．例えば，要件違反の予測結果に対してはより多くの修正を

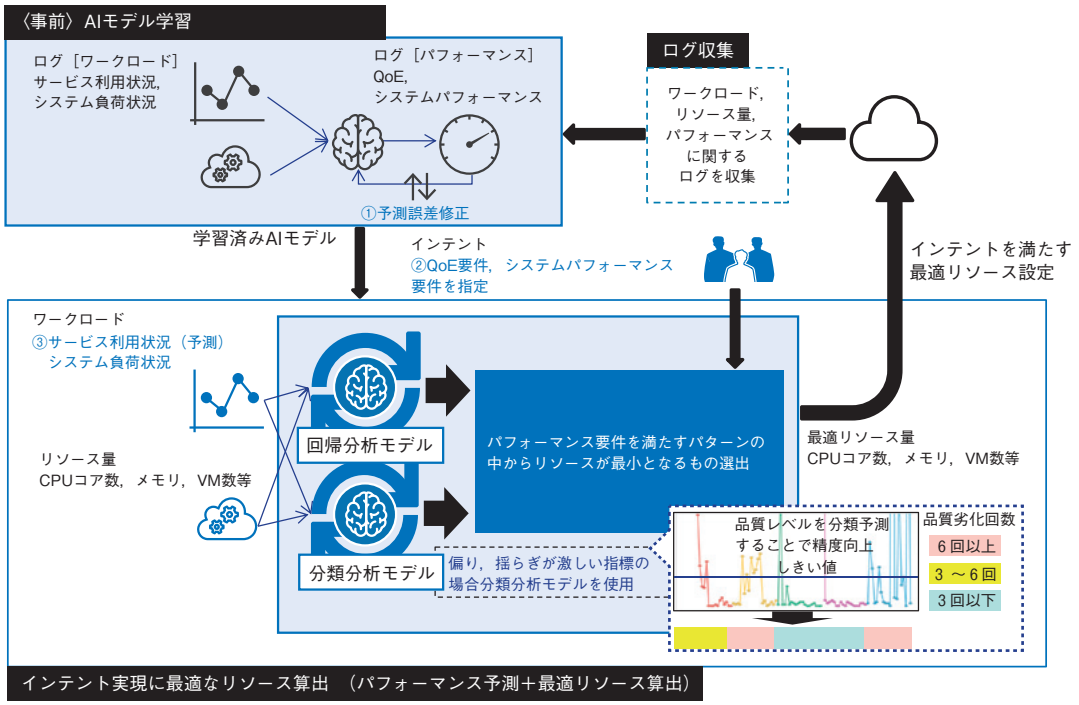


図 1　技術概要

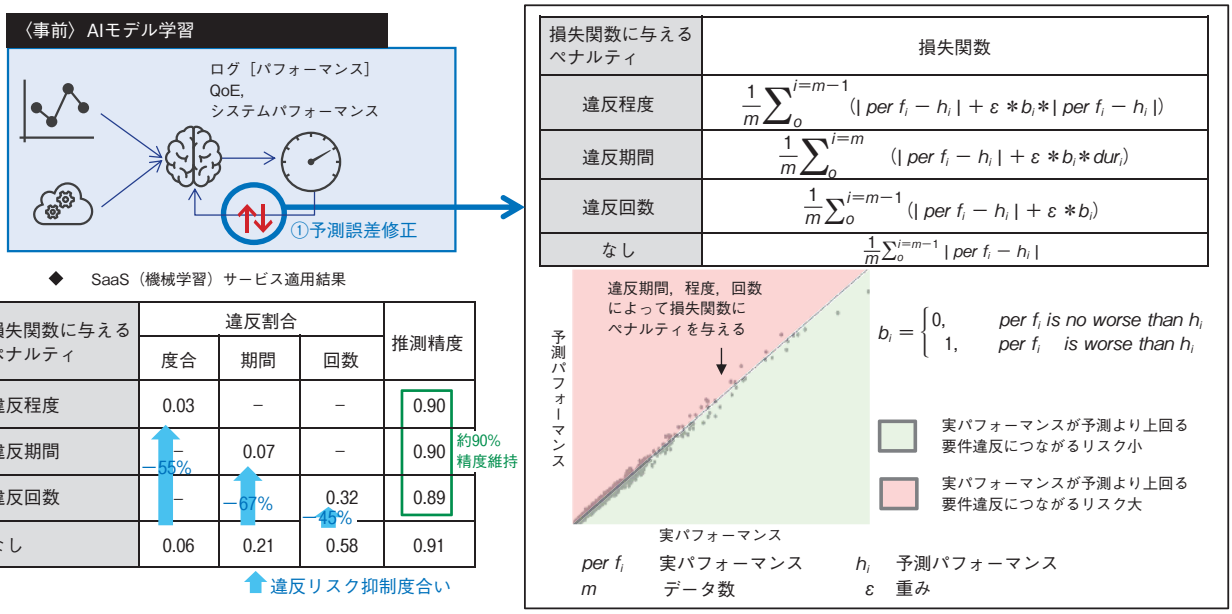


図 2　要件違反リスク抑制機能

実施するなど，損失関数自体も独自のものを用いています（**図 2**）．しかも，単に違反の度合いだけでなく，ネットワークオペレーション分野での性能管理に重要な，違反期間・違反回数に対して損失関数を個別に定義しています．

また，サービス（アプリケーション）とシステム（インフラ），双方のワークロード（図 1 ③）をAIの入力とすることにより，システム負荷状況に加えて，サービス利用状況（予測を含む）に応じたリソース算出が可能であるという優位性も持っています．これにより，サービス提供者の需要予測や外的環境変化の情報に合わせてプロアクティブに制御することが可能となり，特に今回事例のWeb会議サービスのように，いったん開始された会議の収容先サーバをサービス提供しながらスケールアップすることや収容先のサーバ変更が容易ではない場合は，この優

位性が効力を発揮します．会議予約情報を基に，数分～数時間先のサービス利用に合わせて事前に最適なリソースを算出することで，会議開始時に収容先の選択や事前のスケールアウト実施を実現することができるためです．

■技術の適用・応用先

本技術は，サービス導入時のリソース設計フェーズでの利用はもちろん，サービス運用フェーズでの動的なリソース制御にもお使いいただくことができます．また，適応対象サービスも今回紹介するWebサービスのようなSaaS（Software as a Service）提供リソースだけでなく，「クラウドサービスを使って，目的の時間内に機械学習を終了するためにはどの程度のリソースが必要か」といったIaaS（Infrastructure as a Service），インフラ利用時にも使うことができます．さらに将来的には，仮想ネットワークサービスを提供するための仮想ネット―ワーク機器への適用視野に拡張していく予定です．

Web会議サービス向け技術適用について

本技術のWeb会議サービスへの適用においては，Web会議サービスが抱える課題や特徴を踏まえて，AIモデル作成における工夫や，新たな機能追加を実施しています．

■Web会議サービスが抱える課題と課題解決イメージ

Web会議サービスは，前述したような社会情勢の変化により，次のような課題に直面しています．

- サーバリソースおよびミーティングの規模に応じたサーバ割当てにより，サーバリソースの最適化およびコスト最適化を実施したい．
- QoE維持のための技術を具備していることを明確化することで，ユーザのサービスに対するブランド価値を向上させたい．

これらの課題に対して，本技術を適応することにより，**図3**のように課題を解決します．

従来の方法では，リソースを設計する際に技術者の経験やさまざまな検証により，想定される最大のサービス利用量（ユーザ数等）に対して，余裕を持ったサーバリソースをあらかじめ用意するのが一般的です．この場合，利用者が少ない状況では無駄なリソースが発生し，想定を超えた利用者がアクセスした状況では，さまざまな品質劣化が発生します．さらに，Web会議のようなサービスでは，運用中のサーバ増強も困難であり，致命的な状況に陥りかねません．

本技術とクラウドサービスが持つ制御機能とを連携させることで，利用者が少ない状況では最小リソース量によりサービスを運用し，利用者の増加に合わせてサーバをスケールアウトさせることで，この問題を解決することが可能となります．

■Web会議向けパフォーマンス予測AIモデルの作成

Web会議サービスにおいて重要となるQoE指標は，映像や音声の奇麗さや安定性に関する指標である，スループット・ジッタなどを用いることが一般的です．これらの値は常に変動している性質があるため，高い精度で

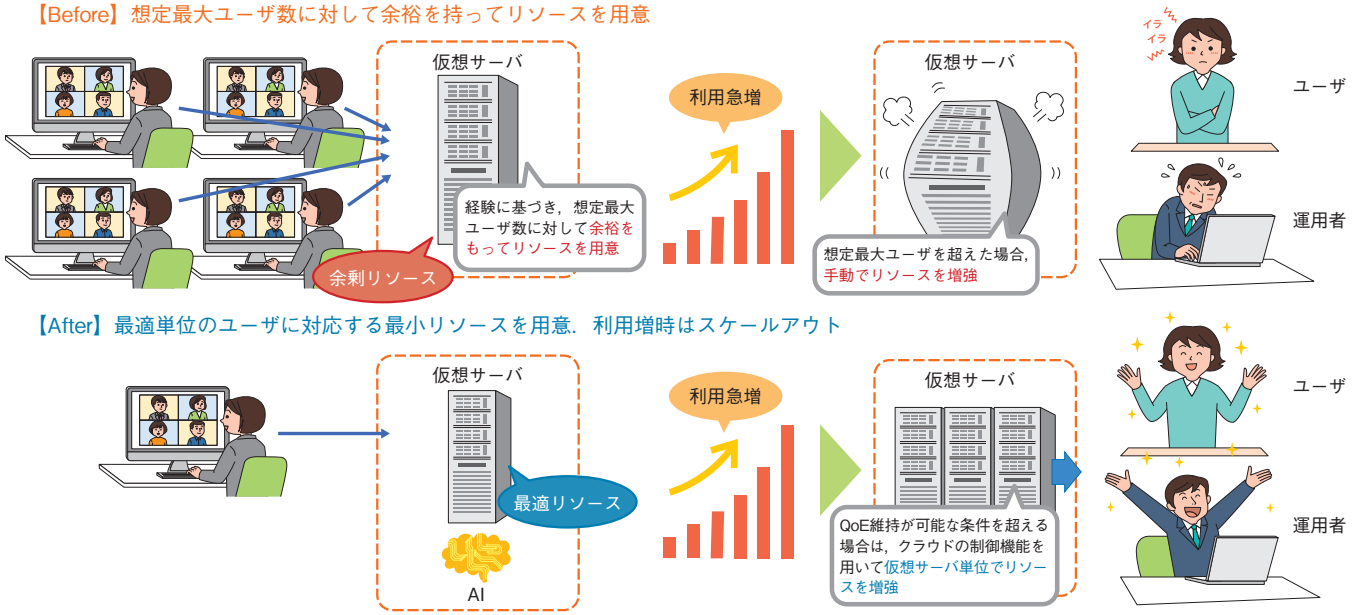
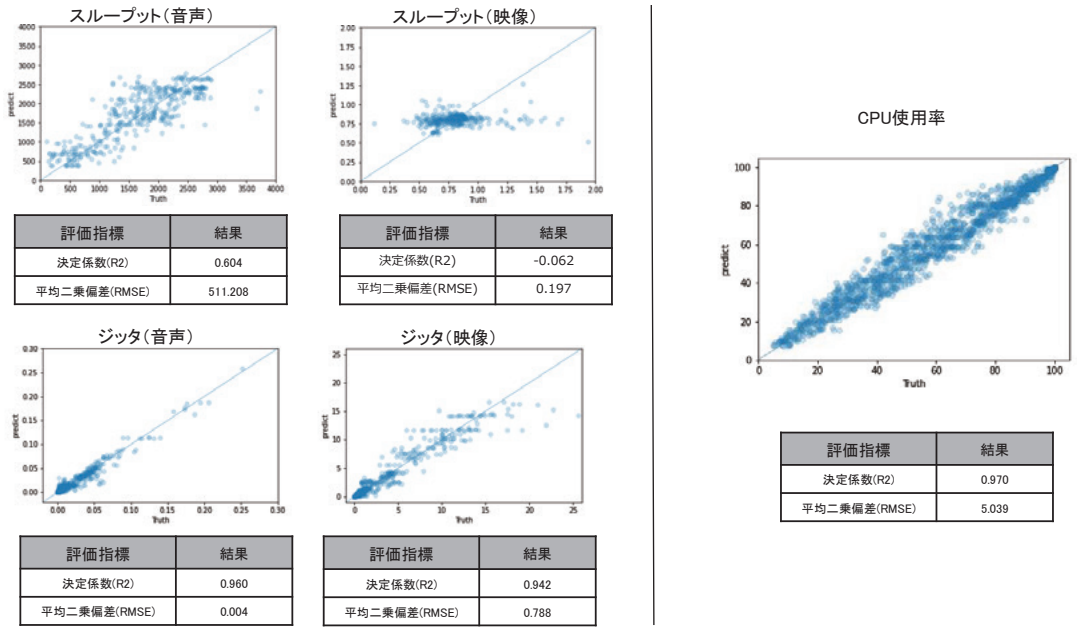


図3　課題解決イメージ

予測することが困難です（**図 4** (a)）．そこで，CPU使用率がある一定以下である状況では，これらの値の劣化が発生しないことに着目し，ワークロード・リソース量からCPU使用率を予測するAIモデル（図 4 (b)）により，QoEを維持するために最適なリソース量の導出を可能としています．

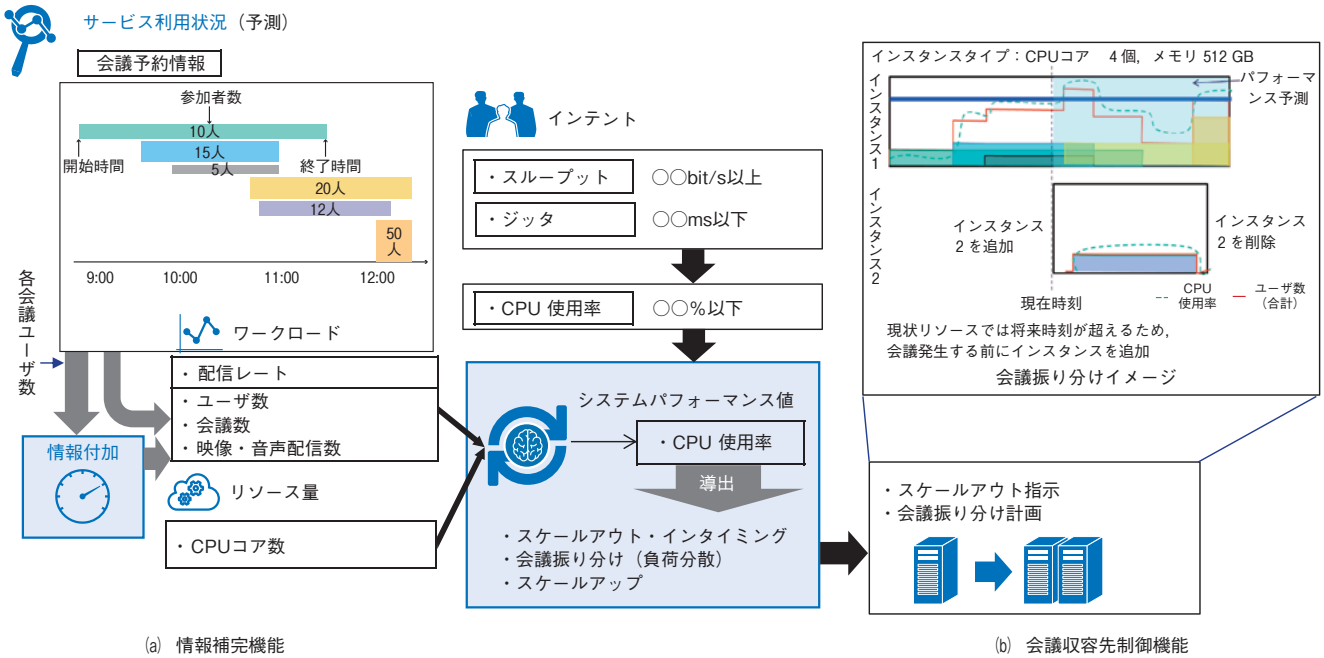
■周辺機能

前述のとおり，本技術をWeb会議サービスに提供するためには，会議が開始される前に，発生後に必要となるリソース量やそれに基づく収容先を決定する機能が必要となります．このため，会議予約情報から実際の会議が発生した際のワークロード情報の一部（映像・音声の配信数）を補完する機能を用意する必要があります（**図 5** (a)）．また，補完されたワークロードに応じて，会議組合せごとの品質指標値を予測し，あらかじめ設定された使用イン



(a) スループット・ジッタ　　(b) CPU使用率

図 4　学習済みAIモデル評価結果



(a) 情報補完機能　　(b) 会議収容先制御機能

図 5　Web会議サービス適用イメージ

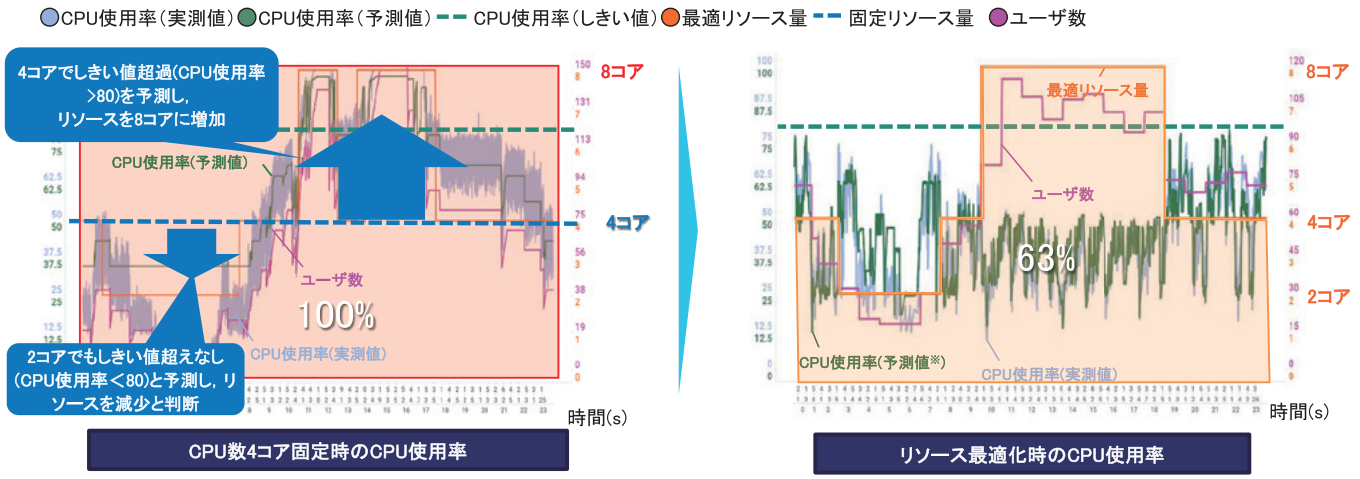


図6　最適リソース算出結果と効果

スタンスタイプ（CPUコア数）への収容可否判断とスケールアウト・インを実施する機能が必要になります（図5(b)）．これらの利用状況（予測）に基づくリソース算出機能を周期的（数分～数時間）に稼動させることで，サービス運用中は常にリソース自動制御を実現することができ，QoEの維持とリソース効率化の両立において，最大の効果を得ることが可能となります．

最適リソース算出結果と効果

実際の商用Web会議サービスの利用傾向を参考に最大100名程度のユーザが利用する想定で利用者のワークロードを生成し，本技術の効果を検証しました．まず，CPUコア数4のインスタンスタイプにおけるCPU使用率を本技術により予測した結果，実際に同数のワークロードを発生させた際の実測値とほぼ同一の結果が得られることを確認しました（**図6**左）．また，これらの予測結果に基づき，リソースが余剰状態になる時間帯をCPUコア数2のインスタンスに変更し，危険域になる時間帯をCPUコア数8のインスタンスに変更することで，CPU使用率をしきい値以下に維持したままサービス提供が可能となり，リソースを余すことなく効率的に利用することが可能となります（図6右）．

これらの結果から，想定されるユーザ利用数に合わせて，短い時間（この場合1時間）ごとに，本技術を用いてリソース制御を実施することで，QoEを維持しながら，サービス利用量に応じてリソースを最適制御した場合，余剰リソースをCPUコア数8で固定的に運用した場合に比べて，約30%のサーバリソース量を削減可能であることが分かりました．

また，このようにサーバリソースの大幅な効率化を実現することで，エネルギー使用量も削減することが可能であり，本技術をさまざまなユースケースに対して適応することで，よりCO_2削減に貢献することができます．

おわりに

本稿では，クラウドサーバリソース最適化技術の概要と優位性について説明し，Web会議サービス適用における工夫と効果について説明しました．

今後はWeb会議サービスユースケースでの商用導入に向けた既存システムとのインタフェース検討や会議振り分け速度向上等の機能改修，さらにはMintentを構成する他の制御技術との連携，IOWN（Innovative Optical and Wireless Network）時代を見据えた他のユースケースへの適用領域拡大の検討を進めるとともに，並行して，「ユーザ・サービス提供者のより曖昧なインテントから定量的なインテントを抽出するための技術」や「ビジネス・サービスレイヤのインテントをリソースレイヤのインテントへの変換する技術」等，新たな技術の創出を進めていきます．

（左から）菊島　宏明/ 呉　超

本稿で記載したような課題をお持ちのサービス提供者の皆様，ぜひお声掛けください．

◆問い合わせ先
NTTアクセスサービスシステム研究所
アクセスオペレーションプロジェクト
TEL　0422-59-4568
E-mail　ibsm-ml @ hco.ntt.co.jp