将来のスマートシティを支える高解像度多カメラ分析基盤

IOWN（Innovative Optical and Wireless Network）時代のスマートシティは，都市のあらゆる情報を価値化しアクセス可能とするCPS（Cyber-physical system）の上に構築されます．本稿では，都市規模のCPSを支える，高解像度・多数のカメラ映像を効率的に処理するAI（人工知能）推論基盤，および基盤上でのAI推論処理の処理量・消費エネルギーの大幅な削減を実現する「イベント駆動型推論」のコンセプトと，その要素技術である「多層推論技術」と「推論リソース共有技術」について紹介します．

三上 啓太　史 旭
井上 規昭　樗林 亮介
松尾 嘉典　山崎 育生
NTTソフトウェアイノベーションセンタ

スマートシティとCPS

皆様がスマートシティという言葉を聞いて思い浮かべるのはどのような都市でしょうか．国土交通省の定義[(1)]によると，スマートシティとは「都市の抱える諸課題に対して，ICT等の新技術を活用しつつ，マネジメント（計画，整備，管理・運営等）が行われ，全体最適化が図られる持続可能な都市または地区」としています．また，野村総合研究所の定義[(2)]では，より具体的に，スマートシティとは「都市内に張り巡らせたセンサー・カメラ，スマートフォン等を通じて環境データ，設備稼働データ・消費者属性・行動データ等の様々なデータを収集・統合してAIで分析し，更に必要に応じて設備・機器などを遠隔制御することで，都市インフラ・施設・運営業務の最適化，企業や生活者の利便性・快適性向上を目指すもの」としています．いずれの定義でも「都市を対象として，ICT技術を活用して管理運用を行い，全体最適を目指す」点は共通しています．こういった，現実世界を対象に情報処理を行い，最適化等を実現する仕組みを表す概念として，サイバーフィジカルシステム（CPS：Cyber-physical system）があります．CPSとは**図1**に示すように，現実世界（フィジカル）の情報を仮想空間（サイバー）に取り込み，コンピューティングによる分析を行ったうえで，その分析結果を再び現実世界にフィードバックすることで，現実世界に最適な作用を及ぼすという，いわば「現実をプログラム可能にするシステム」です．

現実がプログラム可能になると何が起こるでしょうか．それは「現実のソフトウェア化」です．実は現実のソフ

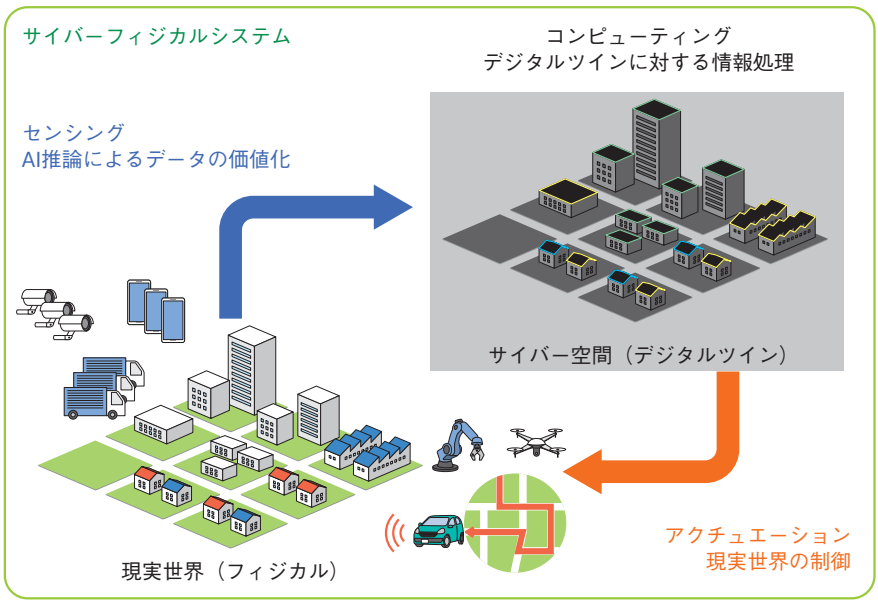

図1　サイバーフィジカルシステム（CPS）

トウェア化はすでにあちこちで起こっています．例えば電話です．かつて電話といえば，それは受話器とダイヤルを備えた箱であり，それが小型化・無線化した携帯電話も，はじめはあくまで遠隔地の音声をつなぐ端末でした．しかし，電話が「ソフトウェア化」したスマートフォンの登場は世界を変えました．物理的なボタンはタッチスクリーン上のアイコンとなり，やがて自由に姿を変えてあらゆる手のひら大のインタフェース（電話，電卓，書籍，カメラ等）を兼ねるようになりました．電子書籍やデジカメも，書籍やカメラがソフトウェア化した例といえるでしょう．このソフトウェア化により，従来の静的で一方通行のインタフェースは動的でインタラクティブなものに変化し，また提示される情報も受け手に応じたパーソナライズやレコメンドが可能となるなど，劇的な変化と利便性・快適性向上が起こりました．

話をスマートシティに戻します．そう，スマートシティというのは，「ソフトウェア化した都市」なのです．これまで都市を構成する要素というのは，コンクリートでできたビルであり，金属製の看板であり，その上で運行される交通機関でした．しかし，スマートシティではビルはBMS（Building Management System）で制御されたスマートビルディングに，看板はデジタルサイネージに，交通機関はコネクティッドカーをはじめとしたスマートモビリティになり，それらはすべてプログラム可能です．都市がソフトウェア化されれば，「いつも混雑する道路」は「交通状況に応じて車線や制限速度が最適化されるスマート道路」になるかもしれませんし，「なかなか来ない路線バス」は「移動したいと思った瞬間に目の前に止まる自動運転タクシー」に置き換わるかもしれません．

都市規模のCPSを構築し，その上でさまざまな都市サービスのソフトウェア化を進めることで，電話がスマートフォンになることで手元に起きたような，劇的な変化と利便性・快適性向上が都市規模で起こることが期待されています．

鍵となるのは AI推論処理における処理量・消費エネルギーの削減

では，スマートシティを支える基盤である都市規模のCPSは，容易に実現可能なのかというと，そんなことはありません．CPSは大きく，センシング，コンピューティング，アクチュエーションの3ステップからなり，それぞれの領域では現在進行形でさまざまな研究が進められています．その中でもNTTソフトウェアイノベーションセンタ（SIC）が課題として取り組んでいるのが，センシングおよびコンピューティングを担うAI（人工知能）推論基盤です．スマートシティにおけるセンシングは，都市の各所に配置されたカメラや各種センサ，都市内のコネクティッドカーやスマートフォンから発生し続ける大量のストリームデータを深層学習による推論（いわゆるAI推論）によって分析し，意味のある情報として価値化することによって行われます．また，価値化された情報を「デジタルツイン」として計算機上に再構築し，現実世界に望む結果をもたらすフィードバックを計算するのにもAI推論は欠かせません．

従来からAI推論は大きな計算リソースを必要とする「重い処理」であり，私たちも「ストリームマージ技術」や「GPUオフローディング技術」といった効率化・高収容化技術に取り組んできました．しかし，IOWN（Innovative Optical and Wireless Network）構想がめざす，「ヒトでは扱いきれない規模の事象をとらえ，ヒトを超える速度で分析・判断できるAIシステムの実現」に向けては，入力データのさらなる高解像度化・高FPS（Frame Per Second）化と，都市規模のカメラ・センサ数への対応が必要となります[(3)]．一般にAI推論の処理量と消費エネルギーは分析対象となるデータ量に比例するため，入力データ量の爆発的な増加は，そのまま処理量と消費エネルギーの爆発的な増加につながります．IOWN時代のスマートシティを支える都市規模のCPSを実現するためには，この処理量・消費エネルギーを持続可能なレベルまで削減する技術が必要となるのです．

イベント駆動型推論のコンセプトと要素技術

前述の処理量・消費エネルギーの削減のために私たちが取り組んでいるのが「イベント駆動型推論」というストリームデータ向けAI推論の実装コンセプトです．イベント駆動型推論の目的は，AI推論に必要な処理量を，「入力データ量依存」から「価値ある情報量依存」に変えていくことにあります（**図2**）．

従来の，すべてのフレームを逐次的に処理する常時処理型のAI推論においては，処理量と消費エネルギーは入力されるデータ量（解像度・FPS・ストリーム数等）に依存します．した

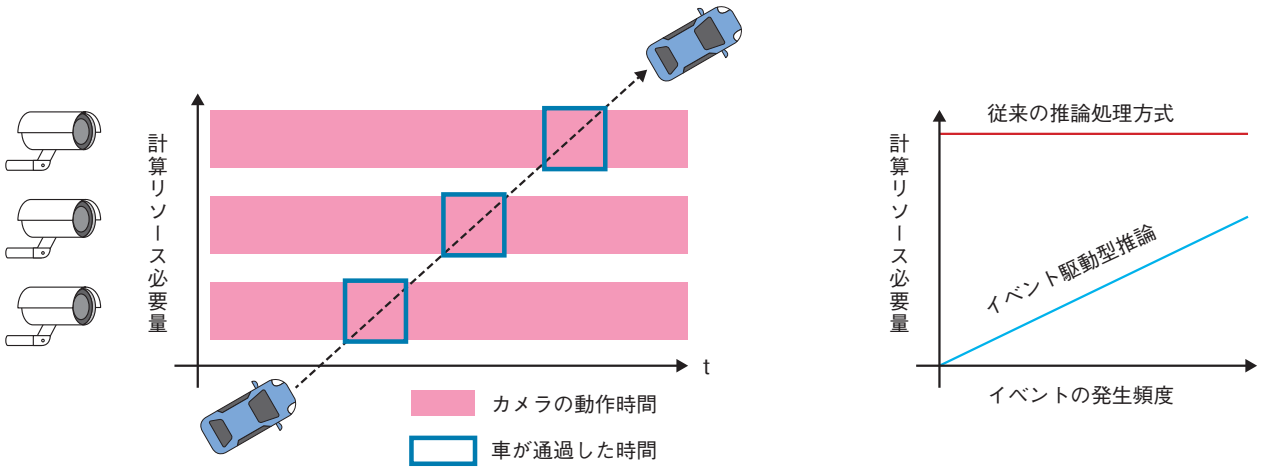

図2　イベント駆動型推論の概観

がって，スマートシティに配置されるカメラやセンサの高解像度化・高FPS化・数量増加は，そのまま処理量と消費エネルギーの増加につながります．これは一見当たり前のように思えますが，例えば人間の認知のメカニズムを考えると，解像度の増加が処理量・消費エネルギーの増加に直結することはありません．それは人間が，普段はなんとなく全体を把握しておき，いざ何かイベント（目の前に急に現れる，音を立てる等）が起きたときにはそちらに注意を払い，対象を詳しく見るというイベント駆動型の認知を行っているためです．この場合，認知に必要な処理量・消費エネルギーは，「注意を払うべき対象の量」すなわちその人にとって「価値ある情報の量」に依存するはずです．AI推論を実装する際も同様の工夫を行うことで，「価値ある情報量依存」の処理量を実現しようというのが「イベント駆動型推論」のコンセプトです．

イベント駆動型推論の実現方式としていくつかのアプローチが考えられます．代表的なアプローチとして，多層推論が挙げられます．多層推論は，同一フレームを複数ステップに分けて分析するもので，次節にて詳しく紹介します．その他のアプローチとして，時間的制御，および空間的制御があります．時間的制御では，時間的に前のフレームの分析結果を基に，以降のフレームの分析パラメータを制御します．例えば，「通常時は5FPSで分析を行い，人物が検知された場合のみ分析の頻度を上げ，15FPSで分析する」といった制御がこれにあたります．空間的制御では，カメラやセンサのトポロジーを活用し，ある入力ソースの分析結果を使って他の入力ソースの分析パラメータを制御します．例えば，「監視エリアの入口のカメラにおいて人物が検知された場合のみ，エリア内のカメラ映像を分析する」といった制御がこれにあたります．

このコンセプトの実現に向けて，SICでは「多層推論技術」および「推論リソース共有技術」に取り組んでいます．

多層推論技術

イベント駆動型推論を実現するためには，何らかの方法でイベントを検知する必要があります．多層推論では前段のイベント検知を行う軽量な前さばきモデルと，後段の本格的な分析を行うAI推論モデルを組み合わせ，イベントが起きたときのみ後段のAI推論モデルを駆動することで系全体での処理量を削減します（**図3**）．前さばきモデルとしては，イベント検知により後段に送るフレームの絞り込みを行うもの，後段と同一のタスクを前段でより軽量・低解像度なモデルで推論し，処理結果の確信度が低い場合のみ後段に送るもの，AI推論を分割してその前半部分を前さばきモデルとし，中間出力を後段に送るものなど，複数のバリエーションが考えられ[3]，ユースケースや利用するハードウェア構成によって最適な構成は変わってきます．

推論リソース共有技術

イベント駆動型推論により処理量が「価値ある情報量依存」になった場合

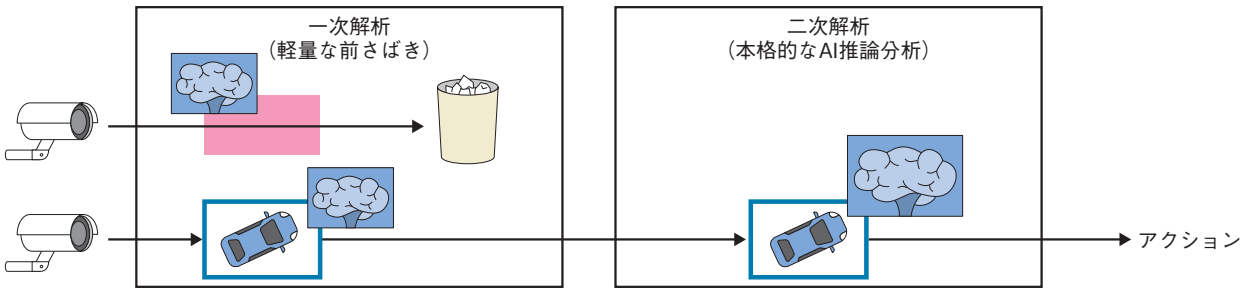

図3　前さばきモデルを用いた多層推論技術の例

に必要となるのが推論リソース共有技術です．従来の常時処理型のAI推論では，必要となるハードウェアリソースは常にほぼ一定であり，その収容設計は容易でした．一方で，イベント駆動型推論では，必要となるハードウェアリソースはイベントに応じて変化します．イベントが集中し負荷が上昇した場合にも対応可能なリソースを用意する必要がありますが，ピーク時に合わせて確保したリソースを普段は遊ばせてしまうのは望ましくありません．

このように，変化する推論の要求に応じて動的にハードウェアリソースを割り当てる技術としてはサービング技術（Triton Inference Server[4]，KServe[5]等）がありますが，一般にサービング技術はオンデマンドなユースケースに特化しています．SICでは，サービング技術をリアルタイムなストリーム処理ユースケース向けに拡張した「推論リソース共有技術」の研究開発に取り組んでいます．推論リソース共有技術を用いることで，リアルタイムなストリーム処理のユースケースにおいても，ストリーム間で推論リソースを共有し，ピークの異なる複数のストリームを束ねて統計多重効果を得ることが可能となります．

今後の展望

本稿では，都市規模のCPSによって実現される，ソフトウェア化された都市（＝スマートシティ）の世界観と，それを支える，高解像度・多数のカメラ映像を効率的に処理するAI推論基盤，およびイベント駆動型推論のコンセプトと，その要素技術である多層推論技術と推論リソース共有技術について紹介しました．

多層推論技術や推論リソース共有技術を利用し，イベント駆動型推論のコンセプトを導入することで，IOWN時代のスマートシティで必要となる膨大なAI推論の処理量・消費エネルギーを大幅に削減することが期待できます．さらに，IOWN に向けた1つひとつの要素技術を積み重ねることで，ヒトを超える速度で分析・判断できるAI システムを実現します．そして，より安全で，誰にでも利用でき，持続可能で，より快適なサービスを創造し，さまざまな社会課題を解決していきます．

■参考文献

(1) https://www8.cao.go.jp/cstp/society5_0/smartcity/index.html
(2) https://www.nri.com/jp/knowledge/glossary/lst/sa/smart_city
(3) 江田・樽林・榎本・史・飯田・羽室：“IOWN時代のAIサービスを支える高効率イベント駆動型推論，”NTT技術ジャーナル，Vol.32，No.12，pp.16-22，2020.
(4) https://developer.nvidia.com/nvidia-triton-inference-server
(5) https://kserve.github.io/website/

（上段左から）三上　啓太／史　旭／井上　規昭
（下段左から）樽林　亮介／松尾　嘉典／山崎　育生

私たちは，現状のデータセンタやクラウドでも動くソフトウェアを開発すると同時に，IOWN を見据えて情報処理基盤に必要な要素技術を創出しています．

◆問い合わせ先
NTTソフトウェアイノベーションセンタ
AI基盤プロジェクト
TEL　0422-59-7262
E-mail　keita.mikami.mp @ hco.ntt.co.jp