GENERALITAT VALENCIANA

SUBVENCIONES DEL PROGRAMA SANTIAGO GRISOLÍA
RESUM DEL PROJECTE

SUBVENCIONES DEL PROGRAMA SANTIAGO GRISOLIA
RESUMEN DEL PROYECTO

GRISOLIAP/2021

A	**DADES DE LA PERSONA INVESTIGADORA SOLICITANTE** ***DATOS DE LA PERSONA INVESTIGADORA SOLICITANTE***			
1r COGNOM / 1º *APELLIDO* ***Manzoni***		**2n COGNOM / 2º *APELLIDO***	**NOM / *NOMBRE*** **Pietro**	**DNI** **X2032102Y**
B	**RESUM DEL PROJECTE D'INVESTIGACIÓ *(indicant el treball a realitzar per la persona a contractar*** ***RESUMEN DEL PROYECTO DE INVESTIGACIÓN (indicando el trabajo a realizar por la persona a contratar***			

Titulo

Optimización de la ejecución de algoritmos de *machine learning* en infraestructuras IoT

Antecedentes

Este proyecto de tesis doctoral se enmarca en la intersección de dos proyectos liderados por miembros del grupo de investigación GRC del departamento DISCA de la UPV. Estos proyectos son **AICO/2020/302** ("FOG-NET: Arquitectura basada en Fog Computing para la optimización de las comunicaciones en entornos IoT"), financiado por la Generalitat Valenciana en la convocatoria de grupos consolidables y el proyecto **H2020-FETPROACT-2020-2** ("SMARTLAGOON - Innovative modelling approaches for predicting Socio-environmental evolution in highly anthropized coastal lagoons"), financiado por la comisión europea en la convocatoria H2020.

El fenómeno de la ***digitalización*** está revolucionando un gran porcentaje de sectores socioeconómicos, gracias a la ***optimización*** de los procesos industriales, generando un tejido productivo más competitivo, pero también proporcionado ***mecanismos novedosos*** para el tratamiento eficiente de aguas, residuos o emergencias con los que afrontar los nuevos desafíos de la globalización o el cambio climático [1, 2]. Uno de los principales impulsores de esta revolución digital es el denominado Internet de las Cosas ***(Internet of Things****, IoT*), donde un gran número de dispositivos interactúan entre sí para ayudar e incluso tomar decisiones de manera autónoma. Dos factores clave sustentan esta revolución: (1) ***los datos*** cuyo análisis puede desvelar patrones ocultos, correlaciones, así como cualquier otro tipo de información potencialmente valiosa, y (2) ***el tiempo de respuesta*** ya que el conocimiento es válido en un marco temporal determinado [3].

La ingente cantidad de datos generada por las infraestructuras IoT así como la velocidad de generación de estos datos requieren de técnicas avanzadas para poder extraer conocimiento. Las técnicas de inteligencia artificial y, en concreto, los algoritmos de aprendizaje máquina o ***Machine Learning*** (ML) están demostrando ser una buena alternativa en este escenario [4]. Estas técnicas permiten diseñar estrategias computacionales (i.e. algoritmos) capaces de aprender automáticamente de los datos de manera agnóstica. Sin embargo, los algoritmos de ML son ***algoritmos complejos*** que requieren ser ejecutados en la última generación de procesadores y sistemas para poder obtener resultados satisfactorios en un tiempo razonable que realmente permita la toma de decisiones [5]. Esto se ha conseguido hasta la fecha mediante la ejecución de estos algoritmos en grandes centros de cálculo ***de forma centralizada***. De esta manera, la información de los sensores se envía en bruto a la nube (*cloud*) donde se procesa de manera eficiente gracias a la disponibilidad de procesadores de última generación como CPUs, GPUs o TPUs [6]. Sin embargo, las infraestructuras IoT cuentan con un gran número de dispositivos emitiendo información de manera periódica y este enfoque centralizado puede ***limitar la escalabilidad*** de las soluciones. Además, lo sensores pueden estar en sitios remotos (p.ej. campos agrícolas, zonas rurales, ramblas, etc.), lugares cerrados, zonas de grandes aglomeraciones o incluso dispositivos móviles, donde la ***conectividad es limitada***, haciendo muy difícil el envío continuo de datos. Finalmente, existen escenarios donde la ***seguridad*** es un factor determinante (p.ej. escenarios médicos) y donde el envío de información a la nube pone en riesgo su integridad [7].

El ***edge o fog computing*** es un paradigma distribuido que consiste en realizar el cómputo en los dispositivos más cercanos a la captura de datos[8]. Esta alternativa esta ofreciendo soluciones a los problemas de escalabilidad, seguridad y conectividad anteriormente citados. Sin embargo, los dispositivos computacionales presentes

actualmente en estos niveles de la red suelen tener poca potencia computacional, normalmente debido a las restricciones de consumo energético establecidas. Incluso con estas limitaciones, el *edge computing* está ofreciendo resultados muy prometedores para cargas de trabajo relativamente poco costosas, relacionadas principalmente con el procesamiento preliminar de los datos como la eliminación de *outliers* o métodos de agregación simples [9].

Las compañías de procesadores como Nvidia, ARM, Intel o Google están ofreciendo ***plataformas hardware de bajo consumo para edge computing*** con procesadores dedicados que incrementa el poder computacional penalizando relativamente poco su consumo energético. Entre estas plataformas destacamos la familia Jetson de Nvidia que incluye un coprocesador gráfico de última generación para la aceleración de código mediante CUDA o el proyecto Coral de Google, que recientemente ha anunciado la evolución de su Edge TPU, que ofrece 4 billones de operaciones por segundo (TOPS), usando 0,5W por TOPS. Sin embargo, la capacidad computacional de estas plataformas sigue siendo mucho menor que las plataformas de servidor, y por tanto, el desarrollo de aplicaciones de IoT no se debería centrar en elegir un ***enfoque cloud o edge***, sino en poder coordinar la infraestructura de manera íntegra y ***transparente al programador***, para que las cargas de trabajo se ejecuten de eficientemente en estos sistemas distribuidos y heterogéneos.

El doctorando seguirá un ***enfoque holístico***, trabajando en **todas** las capas computacionales, desde hardware mediante la evaluación de diferentes configuraciones, tecnologías y dispositivos IoT hasta software, optimizando métodos de *machine learning* en los procesadores emergentes de *edge computing* así como diseñando estrategias de planificación de recursos en el sistema. El plan de tesis doctoral tiene como hipótesis de partida la ***autonomía y ubicuidad*** de las nuevas infraestructuras IoT. Estas infraestructuras están dotadas de capacidades computacionales muy heterogéneas, siendo muy complejo desarrollar aplicaciones avanzadas y, en particular, aplicaciones basadas en algoritmos de *machine learning* que exploten todos sus recursos. Así, el plan de tesis propone el ***diseño de un software de sistemas*** que optimice la ejecución de algoritmos de *machine learning* en infraestructuras IoT de manera transparente al programador.

Se espera que este plan de tesis doctoral tenga un alto impacto científico en forma de publicaciones relevantes en el área pero también que sus resultados sean trasladables al tejido productivo mediante convenios/contratos de transferencia tecnológica con empresas del sector que quieran incorporar la tecnología desarrollada en esta tesis en sus líneas de negocio.

Referencias

[1] Jimeno-Sáez, P., Senent-Aparicio, J., Cecilia, J. M., & Pérez-Sánchez, J. (2020). Using Machine-Learning Algorithms for Eutrophication Modeling: Case Study of Mar Menor Lagoon (Spain). *International Journal of Environmental Research and Public Health*, *17*(4), 1189.

[2] Terroso-Saenz, F., Muñoz, A., & Cecilia, J. M. (2019). QUADRIVEN: A framework for qualitative taxi demand prediction based on time-variant online social network data analysis. *Sensors*, *19*(22), 4882.

[3] Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future generation computer systems*, *29*(7), 1645-1660.

[4] Cecilia, J. M., Timón, I., Soto, J., Santa, J., Pereñíguez, F., & Muñoz, A. (2018). High-Throughput Infrastructure for Advanced ITS Services: A Case Study on Air Pollution Monitoring. *IEEE Transactions on Intelligent Transportation Systems*, *19*(7), 2246-2257.

[5] Cebrian, J. M., Imbernón, B., Soto, J., García, J. M., & Cecilia, J. M. (2020). High-throughput fuzzy clustering on heterogeneous architectures. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2020.01.022

[6] Wang, Y. E., Wei, G. Y., & Brooks, D. (2019). Benchmarking TPU, GPU, and CPU platforms for deep learning. *arXiv preprint arXiv:1907.10701*.

[7] Papadokostaki, K., Mastorakis, G., Panagiotakis, S., Mavromoustakis, C. X., Dobre, C., & Batalla, J. M. (2017). Handling big data in the era of internet of things (IoT). In *Advances in Mobile Cloud Computing and Big Data in the 5G Era* (pp. 3-22). Springer, Cham.

[8] Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), pp 30-39.

[9] Guillén-Navarro, M. A., Martínez-España, R., López, B., & Cecilia, J. M. (2019). A high-performance IoT solution to reduce frost damages in stone fruits. *Concurrency and Computation: Practice and Experience*, e5299.

Descripción del proyecto

El ***objetivo principal*** de este proyecto es optimizar la ejecución de algoritmos de *machine learning* en infraestructuras IoT para habilitar la analítica de datos en tiempo real. Este objetivo general se divide en los siguientes objetivos específicos:

O1.- Análisis, diseño y evaluación de distintas configuraciones de ***infraestructuras IoT***. Este objetivo pretende analizar las principales características de las infraestructuras IoT desde el punto de vista computacional. Se analizarán las principales tecnologías de redes IoT (p.ej. LoRa, NB-IoT, etc.) para determinar sus principales limitaciones (i.e. cuellos de botella) así como diferentes dispositivos de *edge computing* conectados a un servidor heterogéneo (CPU+GPU) mediante diferentes tecnologías (p.ej. Ethernet, WiFi o GPRS).

O2.- Aceleración de algoritmos de ***machine learning*** mediante técnicas de codiseño hardware-software. FASToT seleccionará los algoritmos de *machine learning* más representativos de clasificación, agrupación y regresión para su análisis y aceleración en plataformas de *edge computing* emergentes. Además, se diseñarán técnicas de *machine learning* multiescala adaptadas a la nueva infraestructura distribuida diseñada. Dichas técnicas permitirán el balanceo de carga entre diferentes dispositivos computacionales heterogéneos dependiendo de factores de sobrecarga, consumo energético y/o rendimiento.

O3.- Incremento del poder computacional de los dispositivos ***edge computing*** mediante técnicas de virtualización. FASToT diseñará técnicas de virtualización para dotar de cómputo adicional a los dispositivos computacionales disponibles en el *edge* sin penalizar su consumo energético. El software de sistemas delegará las cargas de trabajo ejecutadas en el *edge* al cloud de manera transparente al programador.

Metodología

Durante el desarrollo de esta tesis doctoral se va a emplear una metodología ágil en cada uno de los campos de investigación del proyecto, que tiene en cuenta la parametrización y el modelado de la arquitectura del sistema IoT, realizando una monitorización y análisis del rendimiento y la energía para los diversos algoritmos de *machine learning* utilizados. Siguiendo la evaluación de los desarrollos parciales como retroalimentación del proyecto para ajustar nuestras implementaciones, así como para corregir errores, se propone reuniones mensuales investigadores expertos en la materia así como con EPOs interesadas en el proyecto. A continuación, se listan las actividades a desarrollar.

Plan de trabajo desglosado en actividades y tareas:

Esta sección presenta el plan de trabajo detallado para conseguir los objetivos planteados en este proyecto. Este plan de trabajo se desglosa en dos niveles de detalle: actividades y tareas. Las actividades planteadas corresponden con cada uno de los objetivos específicos del proyecto.

Actividad 1. Despliegue y evaluación de una infraestructura IoT con dispositivos de edge computing.

Descripción: En esta actividad se desplegará una infraestructura IoT parametrizable con diferentes dispositivos de *edge computing* incluyendo nodos de la familia Jetson de Nvidia y Coral de Google. Estos dispositivos se conectarán con diferentes tecnologías de red a un servidor heterogéneo que actuará como *back-end* e incluirá CPUs multicore y varias GPUs de última generación. Se analizará en profundidad la arquitectura desde el punto de vista de las comunicaciones, consumo energético y rendimiento de todos los dispositivos de la red.

Tareas:
Tarea 1.1 Despliegue de la infraestructura IoT con dispositivos para *edge computing*.
Tarea 1.2 Evaluación de la conectividad (*edge-cloud*) con diferentes tecnologías de red cableadas e inalámbricas.
Tarea 1.3 Evaluación del consumo energético y rendimiento de los nodos *edge computing* y el servidor.

Actividad 2. Análisis, diseño, implementación y aceleración de métodos de *machine learning* en entornos IoT.

Descripción: En esta actividad se explorarán diferentes técnicas algorítmicas de *machine learning* presentes en librerías estándar como Scikit-learn (Python) o RAPIDS (NVIDIA). Estos métodos se evaluarán en plataformas de *edge computing* y en servidores para evaluar las diferencias computacionales entre ambos entornos. Además, se optimizarán estos algoritmos, se diseñarán nuevas versiones y se implementarán estrategias multiescala que permita el análisis de datos en distintos niveles (edge-cloud) y granularidades. Dichas estrategias permitirán el balanceo de carga entre los diferentes dispositivos computacionales heterogéneos dependiendo de factores de sobrecarga, consumo energético y/o rendimiento.

Tareas:

Tarea 2.1 Evaluación de librarías de *machine learning* en entornos edge y servidor.
Tarea 2.2 Análisis, diseño, implementación y aceleración de diferentes algoritmos de *machine* en los entornos IoT desarrollados.
Tarea 2.3 Evaluación del rendimiento y la calidad de los algoritmos desarrollados.
Tarea 2.4 Diseño de estrategias algorítmicas multiescala para la ejecución óptima en entornos IoT heterogéneos.

Actividad 3. Elaboración de estrategias de virtualización para incrementar la capacidad computacional de nodos de bajo consumo.

Descripción: En esta actividad se diseñará un software de sistemas para dotar de cómputo adicional a los dispositivos más próximos a la adquisición de datos intentando penalizar su consumo energético lo menor posible y de manera transparente al programador. Se valorará dos enfoques, (1) la planificación de contenedores (e.g Docker) que contengan algoritmos de *machine learning* mediante la adaptación de herramientas de despliegue de contenedores como Kubernetes y (2) la planificación de código ejecutable en el *edge* al servidor mediante técnicas como rCUDA que virtualizan las GPUs de nodos remotos. Estas técnicas no están preparadas para entornos de bajo consumo como el *edge computing*. De esta manera se evaluarán, propondrán soluciones y mejoras sobre ellas para hacer viable el desarrollo en el marco temporal del proyecto.

Tareas:
Tarea 3.1 Evaluación de Kubernetes para el despliegue de contenedores en el edge.
Tarea 3.2 Evaluación de rCUDA para la ejecución de códigos CUDA de *machine learinng* en el edge.
Tarea 3.3 Diseño de optimizaciones para la optimización de la planificación de recursos en entornos *edge computing*.

Cronograma

Actividades y tareas	Año 1				Año 2				Año 2			
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
Tarea 1.1 Despliegue de la infraestructura IoT con dispositivos para edge computing.												
Tarea 1.2 Evaluación de la conectividad (edge-cloud) con diferentes tecnologías de red cableadas e inalámbricas.												
Tarea 1.3 Evaluación del consumo energético y rendimiento de los nodos edge computing y el servidor.												
Tarea 2.1 Evaluación de librarías de machine learning en entornos edge y servidor.												
Tarea 2.2 Análisis, diseño, implementación y aceleración de diferentes algoritmos de machine en los entornos IoT desarrollados.												
Tarea 2.3 Evaluación del rendimiento y la calidad de los algoritmos desarrollados.												
Tarea 2.4 Diseño de estrategias algorítmicas multiescala para la ejecución óptima en entornos IoT heterogéneos.												
Tarea 3.1 Evaluación de Kubernetes para el despliegue de contenedores en el edge.												
Tarea 3.2 Evaluación de rCUDA para la ejecución de códigos CUDA de machine learinng en el edge.												
Tarea 3.3 Diseño de optimizaciones para la optimización de la planificación de recursos en entornos edge computing.												

Impacto socio-económico, diseminación de resultados y transferencia tecnológica

Sin lugar a dudas, la dupla IoT+IA es un pilar fundamental e indivisible del proceso de digitalización de los diferentes sectores socioeconómicos actuales. Las empresas esperan de este proceso la optimización de su producción y que esto se traslade en beneficios económicos que les permitan ser competitivos en sus respectivos sectores. Este proyecto ofrece la investigación necesaria para hacer factible esta dupla en escenarios reales donde los tiempos de respuestas son cruciales. De esta manera se espera que los resultados de este proyecto sean transferibles a la industria e instituciones españolas en primera instancia, haciéndolas más competitivas y fiables, e internacionales en última estancia para poder abrir mercados internacionales que faciliten la constitución de una spin-off tecnológica que pueda generar puestos de trabajo cualificados. Remarcar que el marco de trabajo de este proyecto se centra en el *edge intelligence* que es una tecnología de vital importancia estratégica y tecnológica para España por su carácter facilitador para el resto de sectores claves nacionales como el sector Turismo (*Smart Turism*), sector agroalimentario (*Smart Agriculture*) o el sector del transporte (*Smart transport*). Esta tecnología permite la apertura de nuevos mercados nacionales e internacionales, gracias a su potencial multidisciplinar, así como la digitalización de contenidos y extracción de conocimiento de manera automática.

Se espera que durante el desarrollo de esta tecnología se genere un número elevado de publicaciones de alto impacto en revistas indexadas (3-4 JCRs Q1-Q2) y congresos internacionales (3-4 Core A-B). Asimismo, se espera que este proyecto sea semilla de un desarrollo tecnológico productivo a nivel científico y empresarial.

La persona sol·licitant
La persona solicitante

firma: PIETRO| MANZONI

Firmado digitalmente por PIETRO|MANZONI
Fecha: 2020.12.10 15:54:50 +01'00'