

# Procesamiento de datos masivos en tiempo real y consumo energético de sistemas paralelos

Javier Balladini<sup>1</sup>, Marina Morán<sup>1</sup>, Claudia Rozas<sup>1</sup>, Rodrigo Cañibano<sup>1</sup>,  
Rafael Zurita<sup>1</sup>, Belén Casanova<sup>1</sup>, Cristina Orlandi<sup>2</sup>,  
Armando De Giusti<sup>3</sup>, Remo Suppi<sup>4</sup>, Dolores Rexachs<sup>4</sup>, Emilio Luque<sup>4</sup>

<sup>1</sup> Facultad de Informática, Universidad Nacional del Comahue  
{javier.balladini, marina, claudia.rozas, rcanibano, rafa}@fi.uncoma.edu.ar;  
{mb.casanova.retamal}@gmail.com

<sup>2</sup> Hospital Francisco Lopez Lima  
orlandi.mariacristina@gmail.com

<sup>3</sup> Instituto de Investigación en Informática LIDI, Universidad Nacional de La Plata  
degiusti@lidi.info.unlp.edu.ar

<sup>4</sup> Departamento de Arquitectura de Computadores y Sistemas Operativos,  
Universidad Autónoma de Barcelona  
{remo.suppi, dolores.rexachs, emilio.luque}@uab.es

## Resumen

Los avances tecnológicos de los sistemas de cómputo paralelo y distribuido permiten el desarrollo de aplicaciones antes impensadas. Una de nuestras líneas de investigación se enfoca en aplicar estas tecnologías a Unidades de Cuidados Intensivos y Unidades de Vigilancia Intermedia. Buscamos mejorar sustancialmente el rendimiento de ellas con el procesamiento en tiempo real de datos masivos generados por el equipamiento médico y otras fuentes. Adicionalmente, trabajamos en la reducción del consumo energético de los sistemas de computación de altas prestaciones, con especial atención en los mecanismos de tolerancia a fallos. Todas nuestras investigaciones se centran en desarrollar metodologías, modelos y soluciones informáticas para colaborar en la resolución de problemas que tengan una alta demanda computacional e impacto social. Los trabajos se desarrollan en colaboración con otras universidades, y un hospital público de Argentina. La formación de recursos humanos en estas líneas está orientada al nivel de grado, maestría y doctoral.

**Palabras claves:** computación de altas

prestaciones, eficiencia energética, big data, aplicaciones para la salud.

## 1. Contexto

Las líneas de investigación aquí presentadas están enmarcadas dentro del proyecto de investigación 04/F013 "Aplicaciones de Cómputo Intensivo con Impacto Social", financiado por la Universidad Nacional del Comahue (UNComa), con inicio el 01/01/2017 y finalización el 31/12/2020, y acreditado por el Ministerio de Educación de Argentina.

Uno de los dos ejes de nuestra investigación, las aplicaciones para la salud, se desarrolla en colaboración con la Unidad de Terapia Intensiva, perteneciente al "Hospital Francisco López Lima" de la ciudad de General Roca, provincia de Río Negro. Integramos el comité de Inteligencia Artificial y Big Data, creado en el presente año, de la Sociedad Argentina de Terapia Intensiva (SATI).

Respecto al eje relacionado con el consumo energético de los sistemas de cómputo paralelo, se desarrolla en colaboración con el Instituto de Investigación en Informática LIDI de la

Universidad Nacional de La Plata (UNLP), y el grupo de investigación HPC4EAS (High Performance Computing for Efficient Applications and Simulation) de la Universidad Autónoma de Barcelona (UAB) de España.

## 2. Introducción

La masificación de las tecnologías de cómputo paralelo hacen que ellas sean cada vez más accesibles, y se pueda pensar en el desarrollo de nuevas aplicaciones. El sector de la salud es un campo que puede ser altamente beneficiado de estos avances tecnológicos. En especial, las Unidades de Cuidados Intensivos (UCIs) pueden mejorar sustancialmente su rendimiento con el uso de Big Data. A su vez, una de las mayores problemáticas que afectan a los sistemas de cómputo paralelo de altas prestaciones (sobre todo a los de grandes dimensiones) es el elevado consumo energético y potencia eléctrica que demandan. Es por esta razón que también enfocamos nuestra investigación en la gestión del consumo energético de estos sistemas, para reducir su impacto económico, medioambiental y social. A continuación se introduce la motivación y problemática que direccionan ambas temáticas de investigación:

### Procesamiento en tiempo real de datos masivos en salud

En los centros de salud, las UCI proveen cuidados continuos y rigurosos para personas adultas que están críticamente enfermas y que pueden beneficiarse de tratamiento, y dan un buen morir a los pacientes irrecuperables. En una UCI, los datos de los pacientes involucran datos clínicos de baja frecuencia, y flujos de datos fisiológicos de alta frecuencia generados por el equipamiento médico (como monitores médicos de signos vitales, respiradores artificiales, etc.). En una UCI típica, los enfermeros completan manualmente datos en formularios, registrando datos clínicos y fisiológicos. Los datos fisiológicos se obtienen por observación de las pantallas del equipamiento médico

(ubicados a los lados de las camas) a intervalos de tiempo (horas o fracciones) que definen los médicos para cada paciente. El equipamiento médico emite alertas cuando hay riesgo en la salud del paciente basándose en las mediciones de ciertos parámetros. Luego, los médicos analizan exhaustivamente los datos de los formularios y dan a los enfermeros las indicaciones de tratamientos a realizar a los pacientes.

Los principales problemas que ocurren en las UCI típicas son:

1. La metodología actual de gestión de la información es propensa a errores humanos. Causas:
  - Podrían cometerse errores en el registro manual, que realizan los enfermeros, de datos observados en los monitores de signos vitales u otro equipamiento médico.
  - El análisis realizado por los médicos, que interrelaciona una gran cantidad de variables, provenientes de enormes volúmenes de información de diferente tipo, origen y soporte (datos fisiológicos, radiografías, análisis de laboratorio, datos de observación del paciente, etc.) es una tarea sumamente compleja. Ella, sin asistencia informática alguna, es difícil de concretar sin cometer errores.Efectos: Posibles diagnósticos erróneos o información inconsistente que implica un esfuerzo adicional del personal para detectar su origen.
2. Detección tardía del deterioro de la salud de los pacientes. Causas: Desde que los datos se toman por los enfermeros hasta que son analizados por los médicos, puede transcurrir un tiempo prolongado. Efectos: Incidencia negativa en la salud de los pacientes, que puede impactar en:
  - Aumento del riesgo de muerte.
  - Aumento de los costos económicos de los tratamientos.
  - Estadías más largas, y por ende la atención de menos pacientes.
3. Pérdida de información de pacientes.

Causas: Los datos observados en el equipamiento médico son registrados en formularios a intervalos de horas (o varios minutos), perdiendo todas las muestras intermedias (por ejemplo, el monitor de signos vitales actualiza la frecuencia cardíaca una vez por segundo).

Efectos:

- Disminución de la precisión de los diagnósticos.
- Imposibilidad de registrar el conjunto completo de todos los datos medidos por el equipamiento médico para uso futuro. Disponer de un gran conjunto de datos históricos permite la extracción de nuevo conocimiento, útil para futuras investigaciones médicas y el descubrimiento de patrones para predicción de patologías. La predicción de patologías, como sepsis (una de las principales causas de muerte en UCIs), permitiría comenzar los tratamientos más temprano, reduciendo el riesgo de muerte, el costo económico de los tratamientos, y las estadías de los pacientes.

4. Ausencia de personal médico especialista: En la mayoría de las UCI del país, no hay médicos intensivistas durante las 24hs. Las Unidades de Vigilancia Intermedia (UVIs) no cuentan con personal médico permanente.

Un gran avance sería disponer de un sistema de cómputo que detecte el deterioro en la salud de los pacientes en tiempo real, basándose en múltiples parámetros de diferentes orígenes. La detección temprana del deterioro permitirá aumentar la efectividad de los tratamientos, y por consiguiente la mejora de la salud de los pacientes, la reducción del costo económico de los tratamientos, y el aumento del número de pacientes atendidos. Ante la ausencia física de médicos en la sala del paciente, el sistema permitiría alertar al personal de enfermería y a médicos intensivistas remotos. Ellos, los médicos intensivistas, podrían observar el estado del paciente de manera remota.

Hay múltiples problemas o desafíos que deben enfrentarse para la construcción de este

sistema, y que creemos viable con la aplicación de técnicas de computación paralela y distribuida, y de los sistemas embebidos. No hay muchos sistemas de este tipo, algunos de ellos se encuentran en etapa experimental inicial y otros ya llevan algunos años de investigación. La información disponible de estos sistemas es normalmente escasa por tratarse mayormente de software privativo. En la bibliografía se encuentran algunos trabajos como [9, 2, 1].

## Consumo energético de sistemas de cómputo paralelo

Mientras el rendimiento de los sistemas de computación de altas prestaciones (HPC, *High Performance Computing*) continúa creciendo, las máquinas aumentan significativamente la cantidad de unidades de procesamiento. Este aumento en el número de componentes hace disminuir la confiabilidad y aumentar el consumo energético de un sistema de cómputo. El consumo energético se ha identificado como uno de los mayores desafíos a enfrentar en la era exaescala [12, 8].

El consumo energético es hoy en día un gran problema. Para dar una idea de la magnitud del mismo, la máquina de mayores prestaciones de la actualidad, la estadounidense Summit, demanda casi 10 MW de potencia, lo mismo que se requiere para abastecer a los hogares de una ciudad con alrededor de 135.000 habitantes (cálculo realizado en base al consumo de un hogar en Argentina). Además del alto impacto económico, la generación de tanta energía podría tener un alto impacto medioambiental, por ejemplo, represas hidroeléctricas que modifican el ecosistema, y social, por ejemplo, la mayor fuente de energía mundial se obtiene del carbón, cuya extracción minera es altamente peligrosa.

La computación ecológica es el estudio y la práctica de la computación ambientalmente sostenible. Ella se ocupa de diferentes aspectos de los sistemas de cómputo: diseño, manufactura, eliminación, y uso. Este último aspecto, el uso ecológico, se refiere al uso de los sistemas de cómputo con conciencia ambiental. Es

posible reducir el consumo de energía de los sistemas de cómputo utilizando diferentes estrategias que deben ser consideradas a nivel del software, y consisten en realizar cambios en la configuración del sistema (escalado dinámico de frecuencia y tensión, hibernación de recursos) o en las aplicaciones (uso adecuado de la jerarquía de memoria, rediseño de algoritmos, planificación de tareas, y asignación de tareas a recursos hardware).

### 3. Líneas de investigación

El eje central de nuestra investigación es desarrollar metodologías, modelos y soluciones informáticas para colaborar en la resolución de problemas que tengan una alta demanda computacional e impacto social en los siguientes campos: aplicaciones para la salud, y el consumo energético de los sistemas de HPC.

#### Aplicaciones para la salud

Esta línea está enfocada en el diseño y desarrollo de un sistema para detección automática y temprana del deterioro de la salud de pacientes en UCIs. Se busca que el sistema, basándose en el conocimiento de expertos y guías clínicas, emita alertas que orienten en la detección de la afección del paciente y en posibles tratamientos a seguir. La construcción del sistema involucra diferentes desafíos tecnológicos relacionados fundamentalmente a: Extracción de datos del equipamiento médico; Procesamiento en tiempo real de grandes volúmenes de datos (Big Data) en tiempo real, generado por el equipamiento médico, especialmente las curvas como el electrocardiograma, la saturación de oxígeno en sangre, entre otros.

#### Consumo energético de sistemas de HPC

Nos centramos en el desarrollo de metodologías, modelos y construcción de software para administrar y gestionar el consumo de energía y prestaciones de sistemas de cómputo paralelo. Nuestro principal objetivo actual es la ges-

ción energética en mecanismos de tolerancia a fallos. La tolerancia a fallos agrega una carga de trabajo significativa al sistema de cómputo, sobre todo en sistemas que tienen enormes cantidades de unidades de procesamiento [7], haciendo necesario gestionar el consumo energético de los distintos mecanismos.

## 4. Resultados y objetivos

#### Aplicaciones para la salud

En [6] presentamos un análisis del estado general de las UTI del hospital Francisco Lopez Lima, y la propuesta del diseño de alto nivel del sistema. En [3] presentamos el diseño de la infraestructura del sistema de procesamiento de reglas y un prototipo, y el desarrollo de un dispositivo embebido adquisidor de datos (actualmente obtiene información de la señal analógica del electrocardiograma desde un monitor médico) que envía datos a la plataforma de procesamiento por medio de una red de WiFi. En [4] proponemos una herramienta para mejorar la prestación de servicios integrados de cuidados intensivos de la salud. Actualmente, estamos trabajando en: desarrollar hardware y software para extraer datos del equipamiento médico, y transmitirlos por WiFi a la plataforma de procesamiento de la información; desarrollar aplicaciones para el procesamiento eficiente de señales; detección de datos fisiológicos erróneos; optimización de la infraestructura de un sistema de Big Data de tiempo real para el procesamiento de reglas médicas; desarrollo de una aplicación para la interacción del sistema con médicos y enfermeros.

#### Consumo energético de sistemas de HPC

Los métodos de tolerancia a fallos tienen fuerte incidencia en el consumo energético de los sistemas de HPC, y resulta de suma importancia conocer, antes de ejecutar una cierta aplicación, el impacto que pueden producir los diferentes métodos y configuraciones del mismo. En [5], presentamos una metodología para

predecir el consumo energético producido por el método de checkpoint coordinado remoto, y en [11] expusimos un análisis de los factores que afectan el consumo energético de operaciones de checkpoint y restart en clusters. En [10] propusimos un modelo para estimar el consumo energético de operaciones de checkpoint y restart, y un método para su construcción. Actualmente, estamos construyendo un simulador para evaluar mecanismos de gestión de tolerancia a fallos que procuren un uso eficiente del cluster, permitiendo maximizar la productividad y minimizar el consumo energético.

## 5. Formación de recursos humanos

El equipo de trabajo local a la Universidad Nacional del Comahue, cuenta con un doctor en la temática (recibido en 2008). En Marzo de 2019 se ha presentado la tesis de grado “Una Infraestructura de Big Data de Tiempo Real para un Sistema de Detección del Deterioro de Pacientes en Terapia Intensiva”, en la Universidad Nacional del Comahue. Se están iniciando dos trabajos de maestría y finalizando una tesis de grado en temas de aplicaciones para la salud. En el corriente año se espera finalizar una tesis doctoral en el tema de consumo energético de sistemas de HPC.

## Referencias

[1] ehcos smarticu, <http://www.ehcos.com/productos/eh-cos-smarticu/> (accedido en marzo de 2020).

[2] Excel medical, <http://excel-medical.com/> (accedido en marzo de 2020).

[3] Javier Balladini, Pablo Bruno, Rafael Zurita, and Cristina Orlandi. An automatic and early detection of the deterioration of patients in intensive and intermediate care units. *Journal of Computer Science and Technology*, 18(03):e25, Dec. 2018.

[4] Javier Balladini, Pablo Bruno, Rafael Zurita, Cristina Orlandi, Remo Suppi, Dolores

Rexachs, and Emilio Luque. A tool for improving the delivery of integrated intensive health care performance. *International Journal of Integrated Care*, 19(4), 07/2019 2019.

[5] Javier Balladini, Marina Morán, Dolores Rexachs, and Emilio Luque. Metodología para predecir el consumo energético de checkpoints en sistemas de hpc. *XX Congreso Argentino de Ciencias de la Computación (CACIC 2014)*, 2014.

[6] Javier Balladini, Claudia Rozas, Emmanuel Frati, Nestor Vicente, and Cristina Orlandi. Big data analytics in intensive care units: challenges and applicability in an argentinian hospital. *Computer Science and Technology (JCST)*, 2015.

[7] Franck Cappello, Al Geist, William Gropp, Sanjay Kale, Bill Kramer, and Marc Snir. Toward exascale resilience: 2014 update. *Supercomputing Frontiers and Innovations*, 1(1), 2014.

[8] Robert Lucas, James Ang, Keren Bergman, Shekhar Borkar, William Carlson, Laura Carrington, George Chiu, Robert Colwell, William Dally, Jack Dongarra, and et al. Doe advanced scientific computing advisory subcommittee (ascac) report: Top ten exascale research challenges. 2 2014.

[9] Carolyn McGregor. Big data in neonatal intensive care. *Computer*, 46(6):54–59, 2013.

[10] M. Morán, J. Balladini, D. Rexachs, and E. Luque. Prediction of energy consumption by checkpoint/restart in hpc. *IEEE Access*, 7:71791–71803, 2019.

[11] Marina Morán, Javier Balladini, Dolores Rexachs, and Emilio Luque. Factores que afectan el consumo energético de operaciones de checkpoint y restart en clusters. *XXIV Congreso Argentino de Ciencias de la Computación (CACIC 2018)*, 2018.

[12] John Shalf, Sudip Dosanjh, and John Morrison. Exascale computing technology challenges. In *Proceedings of the 9th International Conference on High Performance Computing for Computational Science*, VECPAR’10, pages 1–25, Berlin, Heidelberg, 2011. Springer-Verlag.