

Intel® oneAPI, novedades y avances 2024

Escuela Invierno CAPAP-H 2024

CGS

27 de enero de 2024

- “Intel® oneAPI Programming Guide”,
<https://www.intel.com/content/www/us/en/develop/documentation/oneapi-programming-guide/top.html>



Outline

1 Introducción

2 Suite oneAPI

3 Intel Developer Cloud

4 Otros





Introducing

oneAPI

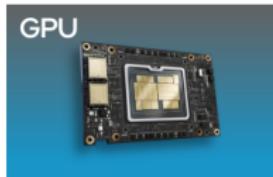
No transistor left behind

Introducción



Introducción

- Alto rendimiento (HPC) solía ser un cuestión **exclusiva** en la gran ciencia
- ... pero está siendo una característica fundamental en otros ámbitos
 - IA, análisis de datos, creación de contenido o gráfico
- Proliferación de arquitecturas : GPUs, FPGAs, ASICs...
- **Programadores lidian con la complejidad de los diferentes modelos de programación en cada tipo de acelerador**



- Creación de la **Unified Acceleration Foundation (UXL)** anunciada en **Linux Foundation Open Source Summit Sep23**
- **Objetivos**
 - ① Unificar ecosistema SW para computación heterogénea
 - ② Modelo de programación abierto y basado en estándares para todos los aceleradores, fomentando el soporte de múltiples arquitecturas y múltiples proveedores

Steering Members

arm

FUJITSU

Google Cloud

Imagination

Qualcomm

intel

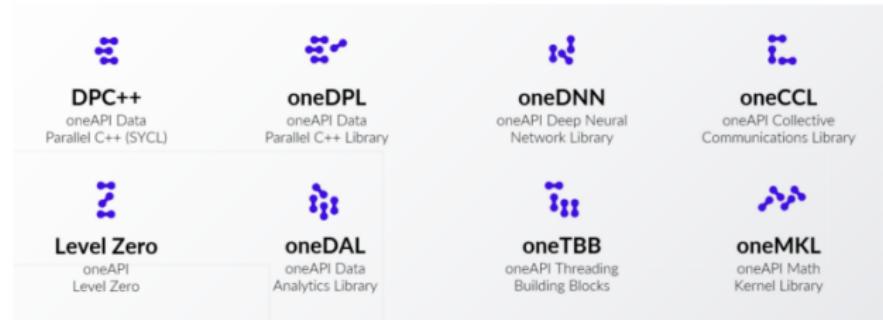
SAMSUNG

UXL FOUNDATION Unified Acceleration



Fundación de UXL

- SYCL como estándar abierto de Khronos y la especificación oneAPI forman la base de los esfuerzos de la fundación UXL
- Colaboración con proveedores de procesadores y desarrolladores de software, alineándose con organismos de estándares como Khronos e ISO C++

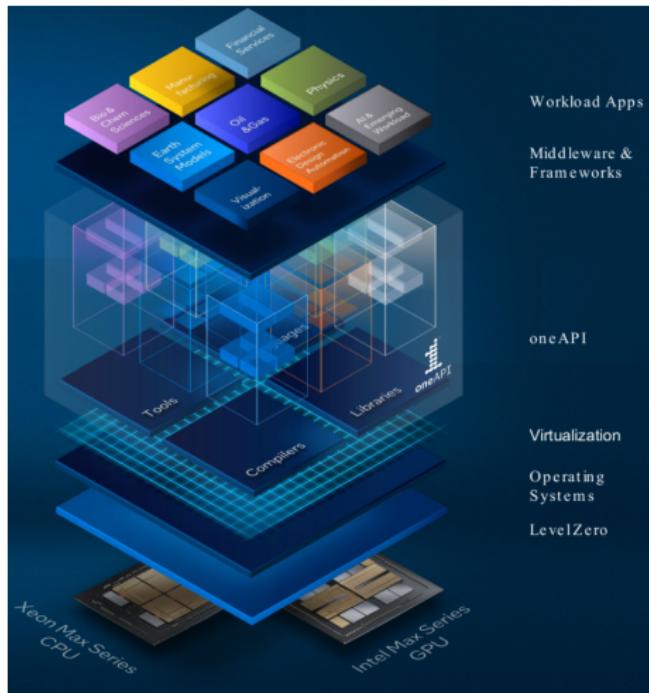


oneAPI elements Intel is donating to UXL



Introducción oneAPI

- Modelo de programación unificado: diversas arquitecturas
- Lenguaje y bibliotecas optimizados
- Rendimiento equivalente lenguaje nativo de alto nivel
- Basado en estándares de la industria y especificaciones abiertas
- Compatible con los modelos de programación HPC existentes



Introducción oneAPI

- Un lenguaje basado en estándares: C++ y SYCL
- Potentes API para acelerar funciones de dominio específico

Soluciones a proveedor único

- Estándar abierto para promover el apoyo de la comunidad y la industria
- Permite la reutilización de código en diferentes arquitecturas y proveedores





Introducing

oneAPI

No transistor left behind

Suite oneAPI



- Un conjunto completo de herramientas de desarrollo testeadas desde CPU a XPU
- Disponible para su instalación en [Intel® oneAPI Toolkits](#)

For most developers

[Intel® oneAPI Base Toolkit](#)

Use Case: Develop performant, data-centric applications across Intel® CPUs, GPUs, and FPGAs with this foundational toolset.

For deep learning inference developers

[Intel® Distribution of OpenVINO™ toolkit \(Powered by oneAPI\)](#)

Use Case: Deploy high-performance inference applications from edge to cloud.

For HPC developers

[Intel® HPC Toolkit](#)

Use Case: Build, analyze, and scale applications across shared- and distributed-memory computing systems.

For visual creators, scientists, and engineers

[Intel® Rendering Toolkit](#)

Use Case: Create high-fidelity, photorealistic experiences that push the boundaries of visualization.

For data scientists and AI developers

[AI Tools](#)

Use Case: Accelerate end-to-end data science and machine learning pipelines using Python* tools and frameworks.

For system engineers

[Intel® System Bring-up Toolkit](#)

Use Case: Strengthen system reliability with hardware and software insight, and optimize power and performance.



oneAPI Toolkits 2024

● Intel oneAPI Base

- Intel® oneAPI DPC++/C++ Compiler
- Intel® DPC++ Compatibility Tool
- Intel® oneAPI DPC++ Library
- Intel® oneAPI Math Kernel Library
- Intel® oneAPI Threading Building Blocks
- Intel® oneAPI Collective Communications Library
- Intel® oneAPI Data Analytics Library
- Intel® oneAPI Deep Neural Networks Library
- Intel® Integrated Performance Primitives



Intel® VTune™ Profiler

Intel® Advisor

Intel® Distribution for GDB*

● Intel® HPC Toolkit

- Intel® Fortran Compiler
- Intel® Fortran Compiler Classic
- Intel® Inspector
- Intel® MPI Library
- Intel® Trace Analyzer and Collector

● AI Tools

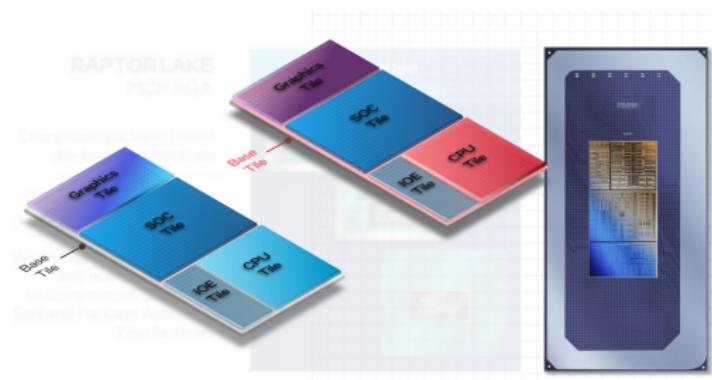
- Intel® Distribution for Python (scikit-learn)
- Intel® Extension for PyTorch*
- Intel® Extension for TensorFlow*
- Intel® Optimization for XGBoost*
- Intel® Optimization of Modin*
- Intel® Neural Compressor
- Intel® AI Reference Models

● Intel Rendering Toolkit

- Intel® Embree
- Intel® Implicit SPMD Program Compiler
- Intel® Open Volume Kernel Library
- Intel® Open Image Denoise
- Intel® OpenSWR
- Intel® OSPRay
- Intel® OSPRay Studio
- Intel® OSPRay for Hydra*
- Intel® Open Path Guiding Library
- Rendering Toolkit Utilities

“AI Everywhere”

- Añadida nuevas características para HW anunciado recientemente
- En el [evento de Dic2023](#) Intel anuncia la gama de productos con “AI solution everywhere”
 - Hw:
 - Procesadores Intel® Xeon 5th **Emerald Rapids** (Intel® AVX512 & AMX Extensions)
 - Procesadores Intel® Core™ Ultra (14th gen): P-cores/E-cores+(low E-cores en SoC tile) + iGPU Arc + NPU

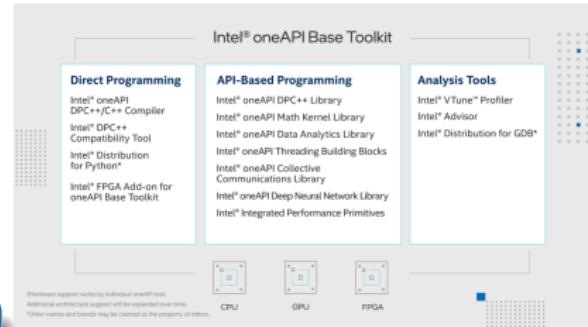


oneAPI Base Toolkit

- Conjunto básico de herramientas y bibliotecas de alto rendimiento
- Compilador de C++ líder con soporte SYCL (computación heterogénea)

Características

- Compilador Data Parallel C++
- Portabilidad con SYCLomatic
- Distribución de Python (librerías optimizadas scikit-learn, NumPy)



Novedades en DPC++

- Mejora el soporte de los aceleradores con más características del SYCL2020 y OpenMP 5.0, 5.1, 5.2
- Estandares C++17 y C17 por defecto aunque tambien soportados C++20, C++17C17, C++14, and C++11/C11
- Mejora el rendimiento en aplicaciones CPU-GPU
 - SYCL/DPC++ ofrece rendimiento equivalente que OpenMP en CPUs

¿Que es DPC++?

- Compilador con soporte de SYCL
- Basado en estandares LLVM



Kernels Paralelos en DPC++



- Expresar paralelismo mediante *kernels* permite que varias instancias de una operación se ejecuten en paralelo
- Útil para descargar la ejecución paralela de un bucle **for-loop** con iteraciones independiente
 - Los *kernels* paralelos se expresan utilizando la función **parallel_for**

for-loop in CPU
application

→ Offload to Accelerator using
parallel_for

```
for (int i=0; i<1024; i++) {  
    a[i] = b[i] + c[i];  
}
```

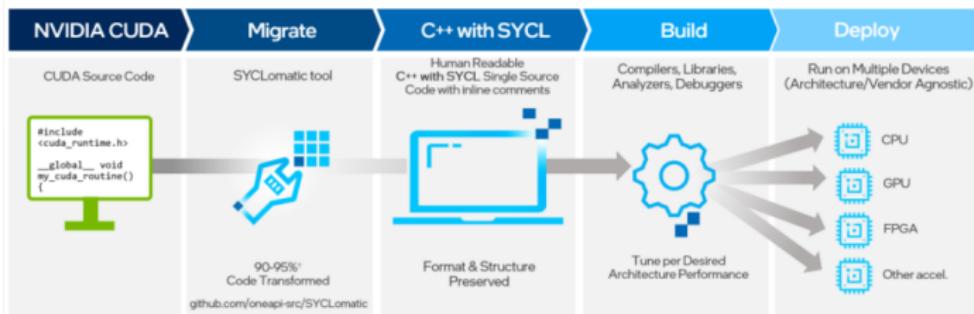
```
h.parallel_for(range<1>(1024), [=] (id<1> i){  
    a[i] = b[i] + c[i];  
});
```



Herramienta de compatibilidad (SYCLomatic)

- Intel® DPC++ Compatibility/**SYCLomatic**
 - Migración para CUDA
 - Soporte del 90-95 % de códigos de CUDA en estudios en varios benchmarks: como Rodinia SHOC, PENNANT
 - Mensajes de Warning son emitidos como salida por línea de comandos
 - Integración con IDE Visual Studio
 - Soporte de CUDA 8.0, 9.x, 10.x, 11.x, 12.0-12.1

CUDA[‡] to SYCL[‡] Code Migration & Development Workflow



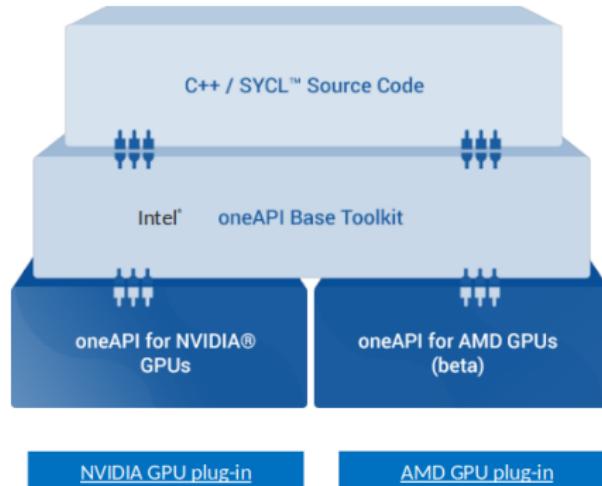
† Intel estimates as of September 2021. Based on measurements on a set of 70 HPC benchmarks and samples, with examples like Rodinia, SHOC, PENNANT. Results may vary.

‡ Other names and brands may be claimed as the property of others. SYCL is a trademark of the Khronos Group Inc.



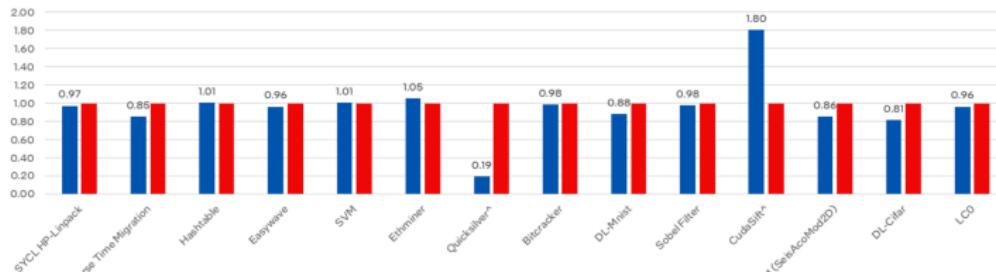
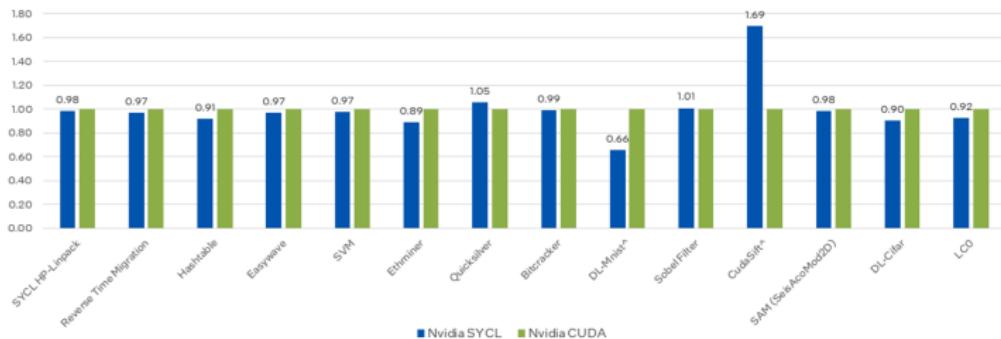
DPC++/SYCL portable

- DPC++ es un proyecto abierto: [fuentes en el github](#) y permite la creación de Toolchain
 - oneAPI para GPUs de NVIDIA mediante plugin desarrollado por codeplay
 - Favorece crear código SYCL y ejecutarlo en GPU NVIDIA compatibles (>sm50)
 - Soporte para GPUs de AMD



DPC++/SYCL portable

- Rendimiento similar en GPUs de NVIDIA (H100) y AMD (Instinct M1250)
 - NVIDIA CUDA vs SYCL
 - AMD HIP vs SYCL

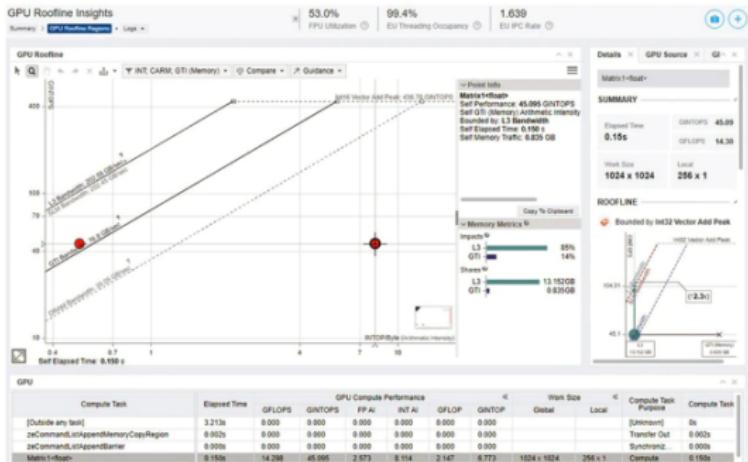


Librerías

- Librerías optimizadas soportadas por compiladores
 - oneDNN usa AMX, AVX-512, VNNI y bfloat16 para aceleración en el proceso de aprendizaje automático
- Bibliotecas optimizadas
 - oneDNN soporta AMX, AVX-512, VNNI y bfloat16 para acelerar el entrenamiento en DL
 - oneMKL mejora la portabilidad y la compatibilidad
 - Integra las optimizaciones vectoriales en RNGs tanto x86-CPUs como GPUs
 - Operaciones optimizados en TF32, FP16, BF16 e INT8.
 - Interfaces para SYCL y C/Fortran OpenMP



- Modelo roofline y rendimiento en CPUs/GPUs
 - Consejos prácticos para generar código en GPUs
 - Recomendaciones de uso CPU vs GPU (uso de jerarquía memoria)
 - Guía de uso en GPUs



- Intel® VTune™ Profiler
 - Herramienta para optimizar el rendimiento de las aplicaciones, el rendimiento del sistema y la configuración del sistema para HPC, nube, IoT, almacenamiento...
 - Análisis de CPU, GPU y FPGA
 - Multi-lenguaje soportado: SYCL, C, C++, C#, Fortran, OpenCL, Python, Google Go...
- Novedades 2024: soporte de perfilado en las nuevas **NPUs** (datos transferidos NPU memoria DDR)
 - Evaluación de tráfico entre CPU y GPU (incluido movimientos USM) para identificar ineficiencias



- Soporte de los estandares a nivel de compiladores:
 - C++17 (defecto), pero también C++20, C++17, C++14 y algunas características de C++23
 - Fortran: cumple estandares de 2018, 2008, 2003... y algunas características de Fortran23
 - Incluye *Coarray* y *DO CONCURRENT* (soporta clausula reducción y “offload” para GPU)
- MPI: soporta completamente las especificaciones MPI-1, MPI-2.2 y MPI-3.1
 - Independencia de red de interconexión



Introducing

oneAPI

No transistor left behind



Intel Developer
Cloud



- Disponible en la URL



- Varias configuraciones que se adaptan a diversas cargas de trabajo
 - Desde capacitación en IA y inferencia
 - ... creación de prototipos y evaluación del último HW utilizar el entorno que mejor se adapte a sus necesidades comerciales
 - ... hasta aplicaciones para FPGA
- Aprenda con tutoriales prácticos
 - Experimente con ejemplos de código del mundo real
 - Evalúe el rendimiento y la aceleración con múltiples configuraciones de hardware.
 - Cree aplicaciones heterogéneas



Hardware disponible

- Se puede testear y evaluar una variedad de máquinas virtuales
 - Sistemas *bare metal*
 - Dispositivos en el Edge
 - Plataformas para entrenamiento de IA
- Entornos para desarrollo
 - Contenedores
 - JupyterLabs
 - Conexión directa por SSH



Instrucciones de acceso

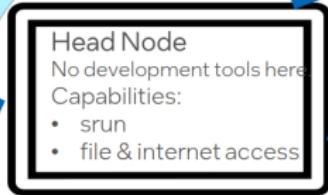
- La documentación y actualizaciones disponible en <https://tinyurl.com/ReadmeIDC> o en el `Readme.md`

Picture it this way

nodes in queues
pvc-shared
and
pvc
are identically configured
same CPUs, same four PC cards
(single tile PVCs – but four of them!)



ssh



Intel
Developer
Cloud
(IDC)

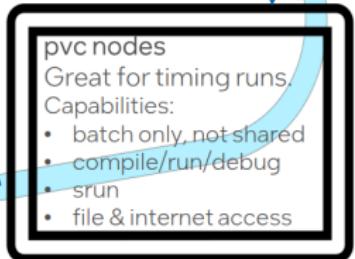
srun



srun



Srun



Training

- JupyterLabs: en el menú **Training and Workshops**
 - Clicar en **LaunchJupyterLab**

The screenshot shows the Intel Developer Cloud interface with the 'Training and Workshops' section selected. The top navigation bar includes links for 'Console', 'Cloud', 'Training', 'Workshops', 'JupyterLabs', 'Dashboard', and 'Logout'. A sidebar on the left features icons for Home, AI, C/C++, SYCL, GPU, and Help. The main content area is titled 'Training and Workshops' and is divided into two sections: 'AI' and 'C++ SYCL'. The 'AI' section contains three items: 'AI Kit XGBoost Predictive Modeling', 'Heterogeneous Programming Using Data Parallel Extension for Numba® for AI and HPC', and 'Machine Learning Using oneAPI'. The 'C++ SYCL' section contains three items: 'Essentials of SYCL', 'Performance, Portability and Productivity', and 'Introduction to GPU Optimization'. Each item has a 'Launch' button with a magnifying glass icon.

- AI**
 - AI Kit XGBoost Predictive Modeling**
Learn predictive modeling with decision trees using Intel® AI Analytics Toolkit
[Launch](#)
 - Heterogeneous Programming Using Data Parallel Extension for Numba® for AI and HPC**
Data Parallel Extension for Numba accelerates Python® code on Intel® XPU
[Launch](#)
 - Machine Learning Using oneAPI**
Intel® AI Analytics Toolkit accelerates data science and analytics with Python®
[Launch](#)
- C++ SYCL**
 - Essentials of SYCL**
Learn to write performant and portable code using oneAPI and SYCL C++
[Launch](#)
 - Performance, Portability and Productivity**
Learn to write performant and portable HPC code for multiple platforms with oneAPI and SYCL C++
[Launch](#)
 - Introduction to GPU Optimization**
Learn GPU optimization techniques using SYCL.
[Launch](#)





Introducing

oneAPI

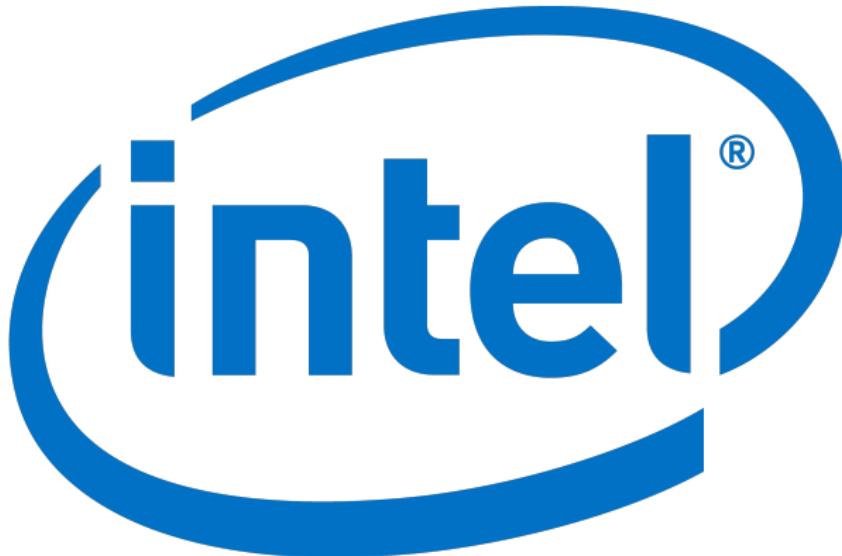
No transistor left behind

Otros

Recursos disponibles

- Iniciativa [oneAPI](#)
- Intel oneAPI Base & HPC Toolkit
- Instrucciones de [acceso del Intel Developer Cloud](#)
- Libro “Data Parallel C++ Programming Accelerated Systems Using C++ and SYCL” (segunda Edición) [disponible online](#)





Software



¡¡¡Gracias!!!



Dirección

Avda. de la industria 4, edif. 1
28108 Alcobendas | Madrid | España

Correo

info@danysoft.com

Teléfono

[+34] 91 663 8683

Sitio Web

www.danysoft.com/intel

